Automated Item Generation by Combining the Non-template and Template-based Approaches to Generate Reading Inference Test Items

By

Eunjin Shin

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Measurement, Evaluation and Data Science

Department of Educational Psychology

University of Alberta

Abstract

Automatic item generation (AIG) is an area of research, where cognitive and psychometric modeling practices are used to create test items with the aid of computer technology. AIG can produce a large number of test items to support the surging demand for test administration. Two general methods are available for producing items using automated processes. The methods vary in their use of templates to structure the content. While the two frameworks could provide significant paradigm shifts to generate test items, the type of test items and the applicability of the items to operational administration are limited. To overcome such limitations, a hybrid AIG framework was created that extends the capacity of template-based AIG with rich natural language processing analyses introduced in the non-template-based AIG systems. The new framework is applied to produce test items in reading comprehension item generation, which is considered a complex and challenging task for previous AIG systems. More specifically, the current method disambiguates an underlying subtopic structure from narrative stories—the Harry Potter series—using topic modelling analysis, a weighted Latent Dirichlet Allocation approach. Then, the disambiguated subtopic information is logically combined and arranged using item models from template-based approaches to generate reading inference-type items. This study has the potential to contribute to the methodology and the current practices of automated item generation by highlighting the importance of integrating two primary components–item models and natural language processing techniques–to generate test items in the previously challenging domain of reading comprehension.

Acknowledgements

I would like to deeply thank my supervisor Dr. Mark J. Gierl for his continuous support and endeavours to help me grow as an independent scholar. I would also like to extend thanks to my committee members, Drs. Hollis Lai, Okan Bulut, Martha White, and Daniel Bolt for their time and support. Moreover, I would like to thank all CRAMErs who have inspired me and encouraged me to constantly grow as a better researcher. Last but not least, this journey would have not been possible without unequivocal love and support from my family. Special thanks go to my partner, Eric Chen, who has been wholeheartedly present throughout this process.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The recent introduction of technological aids in educational assessments has brought a drastic paradigm shift in test item development practices. Unlike the traditional paper-and-pencil test, computer technology has enhanced the full cycle of test administration, delivery, scoring, and feedback (Gierl, Bulut & Zhang, 2018; Susanti, Tokunaga, & Nishikawa, 2020; Zilles, West, Herman, & Bretl, 2019). The technological innovation called computer-based testing (CBT) has produced significant benefits to educators by providing efficient testing practices that can be used in various domains. Such benefits included the drastic increase in test administration frequency, the possibility of incorporating test items using multimedia and multilingual formats, and the possibility of providing examinees with immediate scoring and feedback (Debuse & Lawley, 2016).

Automated item generation (AIG) was introduced to support this transition to CBT. AIG refers to technological innovations that focus on increasing the capacity to produce high-quality and large-quantity test items (Gierl & Lai, 2013). AIG is needed to support the increased demand for a large number of items for test construction in CBT (Kyllonen, 2009). Test construction in CBT requires a pool of large amount of test items that could be used to construct parallel test forms. In constructing parallel test forms, test items of the same level of complexity and target construct are required. However, with traditional item writing processes, a large item bank could not be effectively constructed and managed to sustain the demand. This is largely due to the labour-intensive and limited supervision in the traditional item writing processes. For instance, the traditional item writing process requires subject matter experts (SMEs) to manually accumulate the relevant content to create, evaluate, and validate each test item individually. Hence, creating test items is one of the costliest tasks in educational testing where a single item

for a high-stakes test costs approximately $1,500 - $2,000 USD to produce (Rudner, 2010). Furthermore, SMEs commonly rely on their content expertise and perspectives to solely manage the process, thus, leaving room for subjectivity and error in item creation (Rush, Rankin, & White, 2016). For instance, Masters et al. (2001) indicated that more than 70% of the textbook questions in a nursing item bank were erroneous due to item-writing flaws.

To overcome these limitations, educational researchers introduced AIG as an innovative solution to encourage efficient and effective item writing practices. AIG attempts to replace the human-judgement components in item writing processes, partly or entirely, with highly structured guidelines to reduce the subjectivity and flaws in item development. Re-scaling the unit of item development from individual test items to upper-level cognitive models to generate a large number of items in both an efficient and cost-effective manner. AIG could also bring surprising benefits, such as the possibility of providing immediate feedback, more systematically designed test items, and a significant reduction in the labours of traditional item writing (Alves, Gierl, & Lai, 2010; Gierl & Haladyna, 2012).

Background of the Problem

Previously proposed AIG frameworks were often classified into two categories based on their primary approaches. The two primary AIG approaches are template-based and non-template-based systems. Early studies in AIG focused on using predefined item templates to generate test items using a template-based approach (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003; Gierl & Lai, 2013; 2016; Kyllonen, Pfeiffenberger, Trapani, & Weng, 2009; Singley & Bennett, 2002). Item templates refer to the primary skeleton of test items, which could contain essential components to define and model test items from a parent item (Gierl & Lai, 2020). Template-based approaches extract and manipulate the essential components of the test

item with other plausible alternate values to generate variations of test items. For instance, Gierl and Lai (2013) introduced a three-stage template-based AIG framework. Their system used the cognitive models and item models developed by SMEs to generate test items by identifying the possible combinations of plausible feature sets using computer algorithms. The cognitive and item models act as well-defined guidelines and templates, which allowed SMEs to identify and explicitly present the key information and features from the content. Then, the features are associated and manipulated by computer algorithms based on predefined constraints. The system was widely implemented in various testing domains to successfully generate high-quality test items in large quantities and thereby producing high item acceptance rates for operational test use (Gierl, Zhou, & Alves, 2008). However, the system was still relatively expert-dependent.

Non-template-AIG systems were developed to minimize the human-intervention in item generation processes. The systems depended on various natural language processing techniques and neural systems to generate test items in the reading domain (Aist, 2001; Brown, Frishkoff, & Eskenazi, 2005; Chen, Liou, & Chang, 2006; Devlin, Chang, Lee, & Toutanova, 2018; Mazidi, 2017; Narayan, Simoes, Ma, Craighead, & McDonald, 2020). Early systems focused on identifying and using syntax (e.g., structure of the text) and semantics (e.g., meaning structure of the text) to generate appropriate test items corresponding to given texts. More recently, sequence-to-sequence neural-based systems were introduced to generate test items without any manual feature engineering to identify appropriate items given reading passages and correct responses. The non-template-based AIG frameworks often yielded test items with a drastic reduction in SMEs' roles by directly identifying and extracting the key features from the text to generate corresponding test items. Despite these benefits, non-template-based AIG items were not preferred in the operational test item development. This outcome was largely due to the large

proportions of “unacceptable” test items created by the systems, indicating that the items do not meet the appropriate quality standards required for operational use by SMEs.

Purpose of the Dissertation

The purpose of this dissertation is to address the limitations of previously proposed template-based and non-template-based AIG frameworks. In this study, I introduced an AIG method, which combines the template-based and non-template-based approaches to generate test items with reduced expert-intervention to increase item development capacity. To demonstrate this idea, a new AIG system was created and used to generate inference-type test items in reading assessments with well-known narrative stories. More specifically, the introduced method focuses on employing item models from the template-based approaches with rich natural language processing techniques from non-template-based approaches to successfully generate test items in reading comprehension domain.

Dissertation Outline

A traditional five-chapter format was used to organize and describe the outcomes of this dissertation. The first chapter introduces the grounding problem to situate the motivation and the background of the study. The second chapter introduces and surveys important literature and theoretical backgrounds, which are necessary to understand the methodological rationale for this dissertation. The third chapter describes the analysis procedures and the system architecture of the novel AIG framework. The fourth communicates the results to validate the test item generation capacity of the proposed system demonstrated in the domain of reading comprehension. The fifth chapter discusses the practical implications of the findings, as well as the theoretical and methodological contributions of the study to the literature. The chapter

concludes by acknowledging the limitations of the current study and the suggestion for future research.

Chapter 2: Literature Review

This chapter introduces essential information that is required to understand the item generation framework proposed in this dissertation. First, the history and the emergence of automatic item generation (AIG) are introduced. Then, the two types of AIG are described, which are the template-based and non-template-based AIG frameworks. The review primarily focuses on the overall system design, strengths, and weakness of the two frameworks in their capacity to generate test items. Next, an overview of the natural language processing approach is surveyed. The purpose of this approach is to understand the overall theme and the topic of the passage. The chapter concludes with a summary that provides an overview of the theoretical and technological frameworks, which are essential to understand the next chapters.

Automated Item Generation in Educational Assessments

The introduction and the wide use of computer-based testing (CBT) in education has dramatically changed the education system to provide efficient and innovative assessment practices. Unlike traditional paper-based testing, CBT supports the primary processes of test delivery, administration, scoring, and reporting using technology (Terzis & Economides, 2011). CBT provides a significant improvement in the assessment development and administration cycle. For instance, CBT allows more frequent formative and summative testing for examineess with increased efficiency in test administration. Also, new types of test items can be constructed with rich item materials (e.g., multimedia) and multilingual translation to provide more authentic assessment experiences to examinees (Debuse & Lawley, 2016; Montoya, Egnatovitch, Eckhardt, Goldstein, Goldstein, & Steinberg, 2004). Moreover, using a specific type of CBT, the length and the difficulty of the test could be adapted depending on the examinee's ability level

(e.g., computerized adaptive testing). Hence, examinees no longer need to respond to the excessive number of test items.

One necessary component to support a smooth transition to CBT is the capacity to generate a large number of test items (Kyllonen, 2009). A significantly larger number of items is required to provide the new types of items with rich item materials and to satisfy the increased concerns about the security (e.g., reducing the exposure of existing items to examinees) of CBT. AIG was introduced as a key solution to the problems with a paradigm shift via technological aids in the traditional item writing processes. The traditional item writing process requires intensive and laborious work by the subject matter experts (SMEs), who manually design and evaluate individual test items.

The traditional item writing process by SMEs is an extremely challenging and complex process that is prone to errors. This is because SMEs are given limited guidance to solely manage the process of identifying, organizing, and evaluating the test-relevant content to generate test items. Then, they are expected to interpret and manipulate the acquired content to write, evaluate, and validate the test items individually based on their subject expertise. Hence, the traditional item writing process largely allowed the subjectiveness of the SME's perspectives and understanding to interact with the quality and the difficulty of the test items. Previous studies have identified that experts' judgement on task complexity was not highly associated with the test item difficulty (Hamp-Lyons & Mathias, 1994). Furthermore, Masters et al. (2001) indicated that more than 72% (2,233 out of 2,913) of the multiple-choice test items written in nursing textbooks were found to include item-writing flaws due to SME's subjective judgments.

Because of the challenge and the laborious work, writing test items was often considered one of the costliest processes in educational assessments where a single item for a high-stakes

test cost approximately $1,500 - $2,000 USD to produce (Rudner, 2010). Considering that every item was supposed to be developed and evaluated individually by SMEs, the cost of test item writing linearly increased with the number of test items required. Moreover, item writing was often highly dependent on the SMEs' content expertise and perspectives, thereby, susceptible to the item writer's subjective opinions regarding the content (Schmeiser & Welch, 2006). Hence, given that the transition to CBT requires a significantly larger number of items to be created for valid assessment experiences, the traditional expert-driven item writing processes were no longer compatible with many large-scale assessments (Drasgow & Mattern, 2006).

AIG frameworks support more cost- and time-efficient assessment practices with various advanced methodological approaches. AIG refers to technological innovations focused on increasing the capacity to produce test items (Gierl & Lai, 2013). This approach mainly focused on capturing and containing the necessary information or features in the pre-identified templates to generate test items automatically. Such key item features could vary from parts of the question statements (or stems), list of options and correct answers, to the reading or graphic prompts of the items. The other AIG framework focused on generating items directly without any pre-specified templates (Brown, Frishkoff, & Eskenzai, 2005; Chen, Liou, & Chang, 2006; Du, Shao, & Cardie, 2017; Gao et al., 2018; Mitkov, Le An, & Karamanis; 2006; Narayan et al., 2020). This approach often focused on using various natural language processing techniques to generate the item stem and options directly from the given resources, such as reading passages. In the next sections, two types of AIG frameworks are described. Regardless of the type and the inclusion of templates, AIG frameworks consistently provided surprising benefits that help overcome the limitations of traditional item writing. Such benefits included the possibility of providing immediate feedback, more systematically designed test items, and a significant

reduction in the labour required for manual item writing (Alves, Gierl, & Lai, 2010; Gierl & Haladyna, 2012).

Overview of AIG Frameworks and Methods

Templated-based AIG. Item templates refer to the primary skeleton of test items. The item skeleton could contain essential components to define and model test items from a parent item (Gierl & Lai, 2016). A parent item provides a rudimentary example to identify the essential components of an item. For instance, the commonly used multiple-choice item format contains three primary elements. The elements include, question statement (i.e., stem), list of options (i.e., a correct answer and incorrect options or distractors), and the auxiliary information (e.g., reading passage, figures, or graphs). Then, the key information about the content can be identified by the values and features provided in the stem, options, and auxiliary information. Hence, the intuition behind using the item templates to generate test items relies on identifying these key components and replacing them with logical and plausible values to generate new items (Singley & Bennett, 2002; Kyllonen, Pfeiffenberger, Tranpani, & Weng, 2009; Bejar et al., 2003).

Singley and Bennett (2002) introduced the Test Creating Assistant (TCA) system, which could generate variant items from an item template by manipulating the contents provided as parts of the stem. The system focused on generating alternative items by replacing the manipulatable variables from the item stem with plausible values (e.g., object, building height, and the planet; Figure 1). In Figure 1, each manipulated variable included three specific values as candidates. Then, based on the selected variable values, the answer key can be generated following the answer constraint (e.g., ½ gt^2). The computed answer is placed as one of the keyed or correct options alongside the three plausible-but-incorrect answers or distractors. Considering that three variables were manipulated with three, two, and three candidate values, respectively,

we could expect the provided example item template to generate, 3*2*3 = 18, eighteen similar test items.

Parent Item	A ball is released from rest from the top of a 200m tall building on Earth and falls to the ground. If air resistance is negligible, which of the following is most nearly equal to the distance the ball falls during the first 4s after it is released? A) 40m B) 80m* C) 120m D) 200m
Variables	- Object: A ball, A rock, An iron - Building Height: 200m, 400m - Planet: Earth, Mars, Moon
Answer Constraint	½ gt^2, where g = acceleration due to gravity, t = time

Example Generated Test Items

A rock is released from rest from the top of a 200m tall building on Mars and falls to the ground. If air resistance is negligible, which of the following is most nearly equal to the distance the rock falls during the first 4s after it is released?
Answer key: 30m

An iron is released from rest from the top of a 400m tall building on Mars and falls to the ground. If air resistance is negligible, which of the following is most nearly equal to the distance the iron falls during the first 4s after it is released?
Answer key: 30m

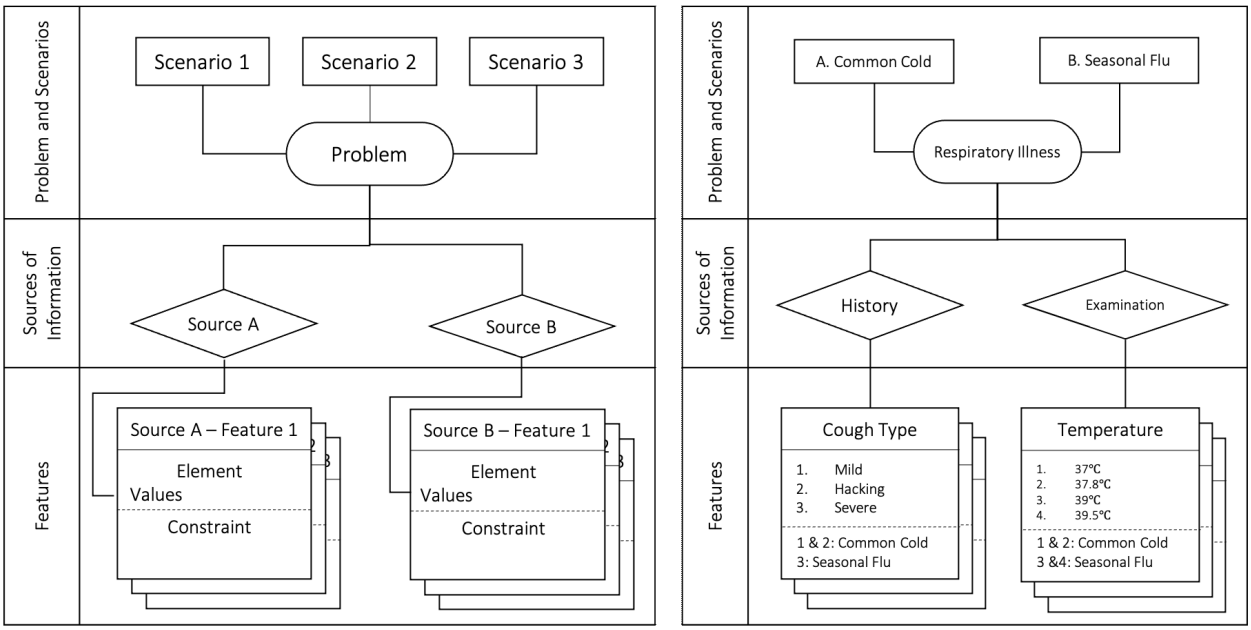
A ball is released from rest from the top of a 400m tall building on the Moon and falls to the ground. If air resistance is negligible, which of the following is most nearly equal to the distance the ball falls during the first 4s after it is released?
Answer key: 13m

Figure 1. Example item-template in Test Creation Assistant of Singley & Bennett (2002).

A more sophisticated template-based AIG framework was introduced by Gierl and Lai (2013). Gierl and Lai's AIG framework consisted of three stages. The three stages included developing cognitive models to identify contents, positioning the contents in an item model, and combining the content. First, cognitive models are developed to explicitly communicate the logic and reasonings used by SMEs to solve specific tasks. Cognitive models identify and convey the mental representations of SMEs in solving the problem given problem-solving scenarios, source of information, and the specific features provided by the source of information (Figure 2).

Second, a corresponding item model is generated using the cognitive model. Gierl, Lai, Hogan, & Matovinovic (2015) indicated and defined the item model as "a template which highlights how the features in an assessment task can be manipulated to produce new items" (p.2). The item model can be considered an equivalent concept of item templates. For instance, the item model for the commonly used 4-option multiple-choice items would consist of three primary components to define an item, which are the stem, elements, and options. Then, the features identified from the cognitive model development in the first stage are directly applied to construct the stem with the variation of key features, or values (Figure 2).

Third, to generate sets of test items from the item model, all possible combination of feature values is assembled following the constraints. This way, a list of test items is generated from one item model with the appropriate feature values manipulated by the carefully designed logics.



Stem	[Feature 1 from Source A] ⋯ [Feature 2 from Source A] ⋯ [Feature 3 from Source A] ⋯ [Feature 1 from Source B] ⋯ [Feature 2 from Source B]
Elements	[Feature 1 from Source A]: values [Feature 2 from Source A]: values [Feature 3 from Source A]: values [Feature 1 from Source B]: values [Feature 2 from Source B]: values
Options	Key 1 Distractor 1 Distractor 2 Distractor 3

Figure 2. Example structures of a cognitive model and an item model of Gierl & Lai's AIG framework (2013).

Unlike the traditional item writing where items were individually written, the template-based AIG framework focused on promoting scalable item development practices. The cognitive models and item models could systematically create a large number of item variants. In other words, items were no longer generated individually, but the underlying models were created to generate large-quanity test items. Gierl and Lai's (2013) AIG framework was applied and

demonstrated in its capacity to produce a large number of high-quality test items in various domains. For instance, Gierl, Zhou, and Alves (2008) described how item models can be constructed in various subject domains, such as language arts, social studies, chemistry, biology, physics, architecture, and mathematics. This indicates the capacity of their AIG framework in generating test items in diverse domains using systematically constructed cognitive and item models.

Benefits and the Limitations of Template-Based AIG. Template-based AIG allows items to be generated following the explicit logic and rules designed by SMEs. Hence, the generated items have a strong theoretical and logical background in terms of how the features are combined and presented to assess the examinees' level of understanding. This allows SMEs to evaluate examinees' misconceptions and misunderstandings more systematically based on their answer choices. Also, because the template-based models focus on developing item-models rather than the individual test items, the generation process was more time- and cost-efficient.

Despite these benefits, one important limitation of template-based AIG is related to the human cost. Template-based AIG heavily relies on the SMEs' capacity to generate structured cognitive models. Gierl and Lai (2013) indicated that training SMEs to generate high-quality cognitive models and to extract relevant information from them to construct item models requires close monitoring with frequent revisions. It is a process that takes time and practice to master.

In addition, template-based AIG is challenging to implement in subject domains where it is not possible to explicitly disambiguate and extract "features" and "values" to represent their contents. For instance, reading comprehension assessments often introduce test items that focus on evaluating the examinees' inferential knowledge. Such items are designed to assess

examinees' ability to evaluate the main idea and the sentiment of the reading passages of varying genre, the length, and types, as shown in Figure 3 (ACT, 2020; The College Board, 2020).

The author describes Henderson's "**Blues in C sharp Minor**" as:

A) Innovative, indulgent, and colorful*
B) Fast-moving, memorable, and eerie
C) Artful, sublime, and unexpectedly upbeat
D) Odd, haunting, and relaxing

The author uses the phrase "**a cathedral of a solo**" (line 85) most likely to create a sense that Berry's solo was:

A) An inticate, awe-inspiring masterpiece*
B) A somber, mournful hymm.
C) A crumbling remnant of Berry's once-great skill.
D) A testament to Calloway's band leadership

Figure 3. Example reading inference items. Retrieved and revised from SAT Evidence-Based Reading and Writing and ACT Reading (p.38; ACT, 2020).

Generating these types of items using cognitive models and item models can be challenging. This is mainly due to the complexity of identifying and capturing the key "features" and "values" related to the answer key and to manipulate them to construct alternative items. For instance, to generate cognitive models to produce inference items as provided in Figure 3, SMEs must identify all text features (e.g., words, phrases, and sentences) used as evidence to support the key answer (e.g., overall sentiment, main topic, or theme). Without any systematic methods to analyze the deep semantic connections between the text features, and the overall themes, main ideas, and the sentiment presented in the story, this can be a challenging and laborious task for SMEs.

Non-template-based AIG. Unlike the template-based approaches, non-template-based AIG aims to generate items directly without any human intervention. Non-template-based AIG creates the essential components to define an item – the stem, options, auxiliary information – directly without any predefined item templates. Thus, non-template-based AIG relies on various natural language processing techniques and neural systems to generate test items (Aist, 2001; Brown, Frishkoff, & Eskenazi, 2005; Chen, Liou, & Chang, 2006; Narayan, Simoes, Ma, Craighead, & McDonald, 2020; Devlin, Chang, Lee, & Toutanova, 2018; Mazidi, 2017). We will review three approaches for non-template-based AIG: syntax-based, semantic-based, and the sequence-to-sequence neural system-based approaches.

The syntax-based approaches generate items by uncovering the underlying syntactic structures of the content. Syntax refers to the set of rules defining how the words and phrases are formed to convey meaning in a sentence or a text. Various syntax analyses in natural language processing techniques were applied in this approach to generate test items. Commonly used techniques included syntax-parsing and part-of-speech tagging (Aist, 2001; Brown, Frishkoff, & Eskenazi, 2005; Chen, Liou, & Chang, 2006; Danon & Last, 2017). For example, Chen et al. (2006) introduced a semi-automated grammar question generation system. Their system focused on identifying parts of the text or the sentences, which could serve as good candidates for grammar questions. This was conducted by matching the regular expressions of the grammatical structure with manually designed patterns to identify the question candidates. The system could also identify a set of plausible but incorrect answers based on their pattern matching. Similarly, Brown et al. (2005) introduced a system that could generate test items assessing examinees' vocabulary. Their system used a large lexical resource, WordNet, to generate six types of vocabulary knowledge questions. The questions were designed to assess examinees' knowledge

of word definition, synonym, antonym, hypernym, hyponym, and cloze questions. The system focused on identifying the correct meaning, or the sense, of the word using the word and its part-of-speech tag information. Once the correct sense was identified for the target word, the system could locate a list of vocabularies that could serve as the keyed answer or the options. While the syntax-based approach could generate test items without item models, the systems were still highly reliant on manually designed patterns and rules about syntactic features identified by human experts.

The semantic-based approaches used the information acquired from the semantic analysis of the content. Unlike syntactic analysis, the semantic analysis focuses on identifying the text features that indicate the meaning interpretation of the content by computer algorithms. Natural language processing techniques, such as the topic modelling, keyword extraction, are used to explicitly model and provide the key information of the text (Flor & Riordan, 2018; Gütl, Lankmayr, Weinhofer, & Höfler, 2011; Mazidi, 2017; Susanti, Iida, & Tokunaga, 2015). Mazidi (2017) introduced a system that could generate test items by identifying and using the primary semantic features from the passages, such as the keywords, main idea sentence, and the summary of the text. His study focused on investigating whether the natural language understanding analysis could help to improve the generated item quality, thus, increase the percentage of acceptable test items. Four methods of analysis were applied to generate an overall understanding of the passages (e.g., topic modelling, terminology extraction, noun phrase extraction, and heading analysis). The resulting items require examinees to summarize the topic with the identified keywords using the stems, such as “Explain what you learned about <keyword> in this passage”, “Explain the relation between brain waves and stages of sleep”, and “Provide a definition for epithelium, and discuss its relation to epithelial tissue”. Then, the

generated items were evaluated by the human raters, in which the items were rated significantly higher than the previously created semantically-based AIG systems in terms of its linguistic and semantic properties.

The sequential neural network approaches were introduced to generate test items automatically. Unlike the previous approaches, neural systems focus on generating test items by disambiguating and learning the primary characteristics of the existing test items. Hence, the system does not distinguish the semantic and syntactic aspects of the passages but rather attempts to learn directly from the test items provided for training. Du, Shao, and Cardie (2017) proposed one of the earliest systems, which attempted to generate a reading comprehension test items using a sequential neural framework. The system focused on mapping the problem statement and the reading passages to generate alternative test items to the problem statement using the recurrent neural networks approach. The system could generate various WH-questions (e.g., "what is one of the largest city centers in the uk?", "when did income inequality fall in the us?", "why do the birds still grow during glacial periods?" ; p.7, Du et al., 2017). Then, the generated items were evaluated by the human annotators, which indicated that the constructed items score high in naturalness (e.g., grammaticality and fluency) and high in difficulty (e.g., difficulty in syntactic divergence and the reasoning).

Similarly, Narayan, Simoes, Ma, Craighead, and McDonald (2020) introduced a next-generation model focused on creating test items that could be directly answered using the information from the reading passage. More specifically, the system used a transformer-based sequence-to-sequence model called BERT (Devlin, Chang, Lee, & Toutanova, 2018). The system was constructed to take the reading passages as an input to generate the most suitable stem or question for the identified answer responses. The model was evaluated on publicly

available question-answering datasets, such as SQuAD (Rajpurkar, Zhang, Lopyrev, & Liang, 2016) and Natural Questions (Kwiatkowski, Palomaki, Rhinehart, Collins, Parikh, Alberti et al., 2019). The human evaluation revealed that the generated factual items were natural and informative. Gao, Bing, Chen, Lyu, & King (2018) introduced a system to generate factual items from reading passages.

Unlike the previous systems, the proposed framework by Gao et al. (2018) could control the item difficulty in the generation process. Their sequence-to-sequence model could take in the sentence from the text with the corresponding answers and a specified item difficulty level (e.g., "easy", "hard") to generate items. The evaluation results indicated the generated items were of high fluency. In Table 1, a list of example test items generated from various systematic approaches in the three non-template AIG categories is provided. The table introduces specific methods, item types, and the example items of the systems.

Table 1

Examples Items Generated from the Previously Proposed Systems

System	Method	Question	Example Items
Chen et al. (2006)	Syntax-based Regular expressions and pattern matching	Grammar question	I intend _____ you that we cannot approve your application. (A) to inform* (B) to informing (C) informing (D) inform
Brown et al. (2005)	Syntax-based Part-of-speech tagging and WordNet	Vocabulary question	Choose the word that best completes the phrase below: The child's misery would move even the most ____ heart. (A) torpid (B) invidious (C) stolid (D) obdurate

Wyse and Piwek (2009)	Syntax-based Syntax parse tree and pattern matching	Reading Factual question	Source text: Emmanuel-Joseph Sieyès trained as a priest and became assistant to a bishop. Q: What did Emmanuel-Joseph Sieyès train as ? Source text: Plate 1 shows an actor dressed as a sans-culotte, carrying the tricolor banner
Heilman and Smith (2009)	Syntax-based Simplified statement extraction using Parse tree	Reading Factual question	Q: What did Prime Minister Vladimir V. Putin return to Moscow to oversee? Who cut short a trip to Siberia? Q: Who was the country's paramount leader? Q: who built his reputation in part on his success at suppressing terrorism?
Agarwal and Mannem (2011)	Syntax-based Syntactic and lexical feature selection, such as the lexical length, token counts	Vocabulary question	An electron having a certain discrete amount of _____ is something like a ball on a staircase. (A) charge (B) energy (C) mass (D) water
Heilman (2011)	Syntax-based Rule-based system with a statistical question ranker	Factual question	Who was deprived of both the knighthood and earldom after taking part in the Jacobite rising of 1715? What is the traditional religion of Japan? Who reorganized the army during the standoff? In 1978, what was awarded the Nobel Prize in Economics? The British evacuated who moved his army to New York City?
Mazidi (2017)	Semantic-based Topic modelling, Term extraction, heading analysis	Factual and Conceptual comprehension question	Explain the relation between brain waves and stages of sleep. Provide a definition for epithelium, and discuss it relation to epithelial tissue
Chali and Hasaon (2015)	Semantic-based Latent Dirichlet Allocation and the Extended String Subsequence Kernel	Factual question	Who designed Apple's first logo? What was replaced by Rob Janoff's "rainbow apple"? What weer conceived to make the logo more accessible?
Mitkov and Ha (2003)	Semantic-based Term exptraction and shallow parsing using WordNet	Factual question	What does a prepositional phrase at the beginning of a sentence constitute? (A) A modifier that accompanies a noun (B) An associated modifier (C) An introductory modifier (D) A misplaced modifier
Aquino et al. (2011)	Semantic-based Information abstraction (e.g., anaphora resolution,	Reading Factual question	Source text: A space conceals the jumping blask. Q: What conceals the jumping blast? Source text: Having take an physics class helped me in Calculus. Q: What helped me in calculus?

	factual statement extraction)		Source text: The customers loved the company's products. Q: What did the customers love?
Yao, Bouma, and Zhang (2012)	Semantic-based Minial recursion semantics text representation	Reading Factual question	Source text: The dog was chased by Bart. Q: Who chased the dog? Source text: John gave the waitress a one-hundred-dollar tip Q: Who gave a one-hundred-dollar top to the waitress?
Du et al. (2017)	Neural-based Attention-based sequence modelling	Factual question	Q:Inflammation is one of the first responses of the immune system to infection. What is one of the first objections of the immune system to infection? Q: From what does photosynthesis get oxygen?
Gao et al. (2018)	Neural-based Encoder-decode model with the long short-term memeory network	Factual question	The electric guitar is often emphasized, used with distortion and other effects, both as a rhythm instrument using repetitive riffs with a varying degree of complexity, and as a solo lead instruction. Q. What is a solo lead instrument? (Hard question) A. The electric guitar Prajñā is the wisdom that is able to extinguish afflictions and bring about bodhi. Q. What is Prajñā is about to bring? (Easy question) A. Bodhi
Narayan et al. (2020)	Neural-based Sequence-to-sequence model & encoder/decode with transformer layer	Factual question	Former Beatle Sir Paul McCartney has topped the Sunday Times rich list of musicians with his £730m fortune. Q. Who is the richest musician in the world? A. Sir Paul McCartney

Benefits and the Limitations of Non-Template-Based AIG. The rapid advancement in non-template-based AIG provides a glimpse into the future of educational test development. Non-template-based AIG can be used to help overcome the limitations of the template-based AIG. They can provide a drastic reduction in SMEs' roles by directly identifying the key features from the reading passages and generate test items directly. In particular, the recent introduction of sequential-neural network-based approaches in AIG required no human input in engineering the features or identifying the set of patterns and rules to generate test items. The semantic-based approach demonstrated how the complex and in-depth meaning structures of the reading passages can be analyzed and extracted to generate test items.

Despite these benefits, non-template-based AIG items are often not preferred in test item development for one primary reason: the generated test items are not ready for operational administration. In other words, the non-template-based approach often produces items that are considered unacceptable compared to the template-based AIG approaches. This is because the generated test items did not satisfy the standards of quality expected for operational testing. Operational testing requires test items to adhere to the quality standard, which defines the relevance and appropriateness of the content, format, and the psychometric properties of the test items (AERA, APA, & NCME, 1999; Drasgow, Luecht, & Bennett, 2006; Lane, Raymond, & Haladyna, 2015). For instance, acceptable test items should present content that is relevant to the target construct and the target difficulty level. Also, they should be free of grammatical errors and presented in a required item format (e.g., constructed-response, selected-response items). Hence, operational test development requires detailed test specifications to communicate information about item quality standards (Downing & Haladyna, 1997).

Using a template-based AIG approach, Gierl, Latifi, et al. (2016), for example, demonstrated how the automatically generated test items could exceed the acceptance rate of the traditionally written items (65% and 52% acceptable for AIG items and SME developed items, respectively). Similarly, Gierl et al. (2015) generated bilingual items (i.e., English and Spanish) in high school science of which more than 90% of the items were considered acceptable, while the rest 10% were considered to need minor revisions when evaluated by two human-raters. Conversely, test items generated from the non-template-based approaches often require heavy editing, filtering, and revisions to extract meaningful and informative test items (Zhang, 2019).

Summary of AIG Methods Review. This review of the two types of AIG indicates the strengths and weaknesses of both approaches. To summarize, the template-based approaches

could ensure the generation of theoretically supported test item generation using cognitive and item models. Thus, the generated items were often considered acceptable and informative for operational testing. However, the intensive labours and amount of expertise required by SMEs to generate high-quality cognitive and item models is challenging for certain subject domains (e.g., reading comprehension). Conversely, non-template-based approaches could demonstrate the potential of significantly reducing human intervention in item generation with intensive natural language processing and machine learning algorithms. While these approaches could generate items directly from the source text and passages, the approaches often suffered from generating fewer proportions of acceptable test items. Still, the non-template-base approach could convey text features, which were more difficult to extract in the template-based approaches, such as the overall theme, main idea, and the sentiment, to generate test items.

Hence, the next chapter introduces a systematic framework that extends the template-based AIG frameworks with techniques used in non-templated based item generation approach. The introduced methods could generate test items that evaluate examinees' overall understanding of the reading passage using a topic modelling algorithm with item models. I turn to this method next.

Natural Language Processing in AIG

Topic Modelling in Reading AIG. Locating the key features for item development from the text, such as the topic structure, has been identified as integral information to create higher-level test items in reading (Mazidi, 2017). Locating documents based on their common topics is a tedious manual task. For example, having SMEs understand, read, and inspect thousands of articles to identify the key ideas and main topics is a complex and time-consuming task. By way of contrast, topic modelling is a machine learning- and natural language processing-based system

that can automatically uncover hidden topics from numerous documents. Therefore, topic modelling can provide methods to automatically organize, understand, search, and summarize large text data without manual human labour (Blei, Ng, & Jordan, 2003).

Topic models refer to statistical models that focus on uncovering the latent structure of the text using the observed word information. Topic models attempt to identify the hidden topic structure from the text. Latent semantic analysis (or LSA) is one of the early attempts at systematically discovering the topic structure from a large corpus (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA attempts to identify the higher-order structural associations among the words in the document to efficiently retrieve documents of similar topics. In LSA, documents are represented by the count information of the vocabularies presented in the text. Then, they are decomposed into two major components, each representing how the documents contain various topics (i.e. document-topic structures) and how the words contribute to defining various topics (i.e. topic-word structures). This can be conducted by various dimensionality reduction algorithms, such as singular value decomposition. The extracted information could uncover the topic information of a large corpus. For instance, given the matrix X with the element (i, j) represent the occurrent of word i in document j, we could decompose the given matrix X to the orthogonal matrices of U, V and the diagnogal matrix of $\sum$ using singual value decomposition or SVD (Equation 1).

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{i,1} \\ \vdots & \ddots & \vdots \\ x_{1,j} & \cdots & x_{i,j} \end{bmatrix}, \qquad (1)$$

$$X = U\Sigma V^{T}.$$

The orthogonal matrix U represents the word-topic matrix, while the V^{T} represents the topic-document matrix. Oftentimes, the SVD is truncated containing only the largest t entities in the

singular value matrix, ∑. Then, t represents the number of topics identified as the results of LSA. (Equation 2, Figure 4).

$$X \approx X_t = U_t \Sigma_t V_t^T. \quad (2)$$

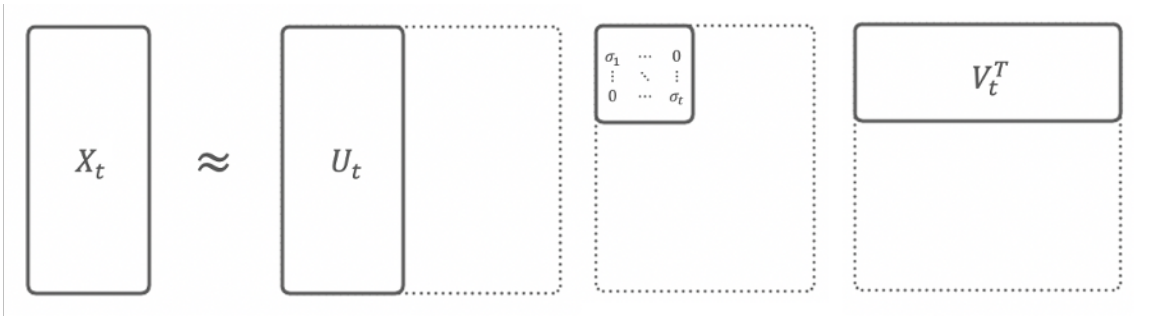


Figure 4. A conceptual representation of SVD in Latent Semantic Analysis.

Probablistic Latent Semantic Analysis (PLSA). Probabilistic Latent Semantic Analysis (PLSA) was introduced to provide a more efficient representation with more interpretable topic outcomes overcoming the limitations of LSA (Hofmann, 1999). PLSA could provide a better statistical foundation to for the previous model by introducing a statisticl model called aspect model as its foundational idea (Hofmann, 2001). The aspect model is a latent variable model, in which the observed variables (i.e., words) are associated with some type of class varible (i.e., topic) in a co-occurrence data (i.e., document). The co-occurrence data is represented using a generative model with the probabilities of the observed and the class variables. In PLSA, a document is represented with two probability distributions, replacing the document-topic and word-topic matrices in LSA.

These probabilities represented a probability of selecting a document D, $P(D)$, a topic distribution given a document, $P(Z|D)$, and the probability of words given the topic $P(W|Z)$. The formal representation of observing given documents with words as the joint distribution can be represented as in Equation 3 and Figure 5.

$$P(D, W) = P(D) \sum_{Z} P(Z|D)P(W|Z) \quad (3)$$

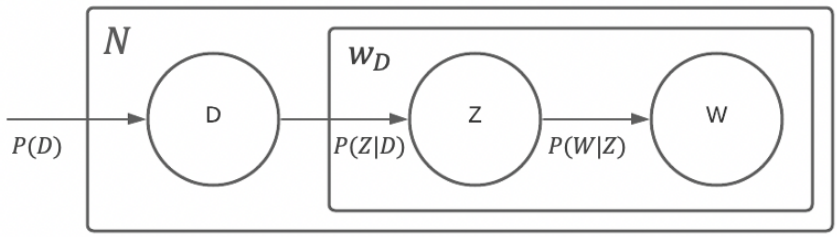


Figure 5. A graph model representation of Probablistic Latent Semantic Analysis (PLSA).

A conceptual representation of the generative topic modelling framework is presented in Figure 6. In Figure 6, Topic 1, 2, and 3 are represented with different types of words, and their distributions, P(W|Z). Also, the common words that are shared between the topics, such as "classroom" and "student", are associated with different weights to define each topic. Then, two documents – document 1 and 2 – are generated by randomly sampling the words from topic 1 and 2, and topic 1 and 3, respectively. Hence, each document was represented with different mixtures of topics, P(Z|D).

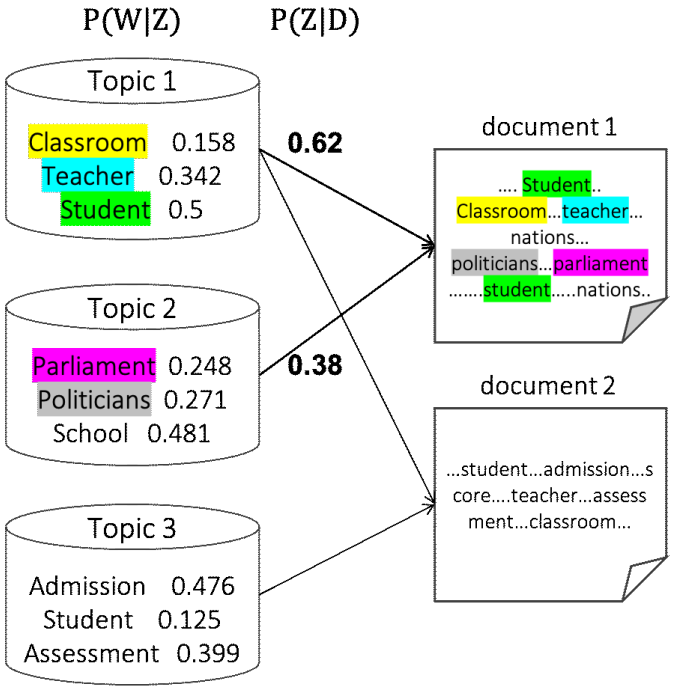


Figure 6. A conceptual representation of the generative topic modelling framework.

The parameters in PLSA are often estimated using the expectation maxmimization (EM) algorithm except the document probability, $P(D)$ is directly estimated from the corpus. The EM algorithm is a general algorithm for the maximum-likelihood estimation (MLE) which iteratively updates the parameters using the E and the M step. The EM algorithm is more suitable for the cases where the data is incomplete or contains latent variables, such as topics, in our case. More specifically, the EM algorithms iteratively introduces the expectation step (or E-step) and the maximization step (or M-step) until the likelihood converges and the best set of parameters are estimated.

While PLSA could provide a good statistical foundation to systematically represent the document generation processes while considering its latent topic structures, it often encountered generalizability issues when provided an unseen document to identify its topic structures. Also, the number of parameters linearly increased with the number of documents to model the topic structure from, which often resulted in overfitting issues. To overcome such limitations, a Bayesian extension of the model was introduced, called the Latent Dirichlet Allocation (or LDA; Blei, Ng, & Jordan, 2003).

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) is a generative probabilistic topic modelling algorithm, in which a document is assumed to consist of a mixture of several topics. These topic structures are referred to as the latent structure of a document, which can be identified based on sets of vocabularies that frequently occur together. To discover the topic structures by locating groups of vocabularies that tend to appear frequently together, LDA also utilizes the two major distributions, word-topic, $P(W|Z)$, and document-topic distributions, $P(Z|D)$, to mimic a document generation process. Unlike PLSA,

LDA introduces Dirichlet priors to the two major distributions to increase its generalizability to discover topic structures in unseen documents. In PLSA, probability of selecting a certain document was directly estimated from the corpus. However, because of the lack of parameters for $P(D)$ with fixed values, it is difficult to assign probability to an unseen document. Moreover, the number of parameters for $P(Z|D)$ would linearly increase with more document introduced to the model, in turn, increasing the concerns for overfitting.

Introducing Dirichlet priors helps to prevent such problems and is a natural choice considering the distributions of the hidden variables in topic models. Because the probabilities of topic given the document and the word given topic are both drawn from multinomial distributions. Thus, Dirichlet distribution, which is a conjugate prior distribution of the multinomial distribution, can be used as prior distribution in modelling the document distribution, $Dir(\alpha)$, and the topic distribution $Dir(\beta)$.

This process can also be described using a graphical representation of LDA provided in Figure 7. For example, topic-word distribution is (ϕ) drawn from a Dirichlet distribution with a hyper-parameter (β). This can be noted as $P(\phi|\beta)$. For each document, documents-topic distributions are drawn from another Dirichlet distribution with a parameter (α), which can be denoted as $P(\theta|\alpha)$. For each N word, a word-topic distribution is chosen as $P(Z|\theta)$ and a word is generated from the word-topic and topic-document assignment, as $P(W|\varphi, Z)$. We then identify a join distribution of a document-topic proportion (θ), word-topic distributions (Z), and the number of words (N) as in Equation 4.

$$P(\theta, \phi, Z, W|\alpha, \beta) = P(\theta|\alpha) \prod\nolimits_{n=1}^{N} P(Z|\theta)P(W|\phi, Z), \tag{4}$$

More specifically, the LDA model assumes that we have M documents, where the M-th document consisted of N_m vocabulary in the document. Then, the topic distribution of the M-th

document is represented as θ_m. This topic distribution, θ_m, is modelled from a Dirichlet distribution with the parameter, α. In terms of the vocabularies, given a fixed number of words, V, and a pre-defined number of topics, K, the vocabulary-topic distribution is represented as $\phi_{k,v}$. This process is iteratively conducted for M number of documents, respectively. Considering that we could only observe the W from the document, it is necessary to estimate the topic-word distribution ϕ_k and the document-topic distribution θ_m efficiently.

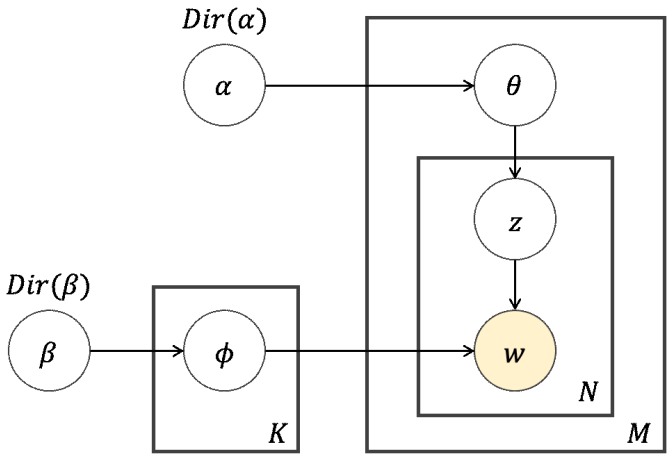


Figure 7. A graphical representation of latent Dirichlet allocation.

Gibbs Sampling. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method which is often used with LDA to effectively extract topic structure from a large corpus. Instead of directly learning the parameters for the topic-word distribution ϕ_k and the document-topic distribution θ_m, LDA uses Gibbs sampling to iteratively learn the topic assignments. Hence, we learn the probability of the word n in the document m to belong to topic k, given the topic assignments of all other tokens, $z_{-(m,n)}$, and the two Dirichlet parameters, α and β. Given that $N_{(n,m,k)}$ represents the number of words n in document m that is assigned to topic k, $N_{(n,m,k)}^{-(n,m)}$ represented the count value when the contribution of the word $v_{(n,m)}$ is excluded (Equation 5).

$$P\left(z_{n,m} = k \middle| z_{-(n,m)}, \mathbf{x}, d, \alpha, \beta\right) \propto \frac{\left(N_{(n,\cdot,k)}^{-(n,m)} + \beta\right)}{N_{(\cdot,\cdot,k)}^{-(n,m)} + V\beta} \cdot \frac{\left(N_{(\cdot,m,k)}^{-(n,m)} + \alpha\right)}{N_{(\cdot,m,\cdot)}^{-(n,m)} + K\alpha}, \text{ where} \quad (5)$$

K: the total number of topics

V: the total number of vocabularies

α: the Dirichlet parameter setting the topic distribution for the documents

β: the Dirichlet parameter setting the topic distribution for the words.

Using this algorithm, we could repeatedly update the topic assignment for each word in each document up to the predefined number of iterations. Once, the topic update is completed, we could integrate out the topic-word distribution ϕ_k and the document-topic distribution θ_m from $z_{n,m}$ as in Equations 6 and 7.

$$\theta_{m,k} = \frac{N_{(\cdot,m,k)} + \alpha}{N_{(\cdot,m,\cdot)} + K\alpha} \quad (6)$$

$$\phi_{k,v} = \frac{N_{(n,\cdot,k)} + \beta}{N_{(\cdot,\cdot,k)} + V\beta} \quad (7)$$

Topic model evaluation. Evaluation of topic model results can be a complicated process, especially for unlabeled text data. When the topic structure of the documents was not pre-identified or labeled by human experts, it becomes more challenging to provides an objective assessment of the performance (e.g., accuracy) of the topic model system. However, manually labeling the topics of a large corpus is not a trivial task. Instead, several approaches were proposed and adopted in the previous topic modelling literature to effectively evaluate unlabeled documents regarding their topic structures (Newman, Lau, Grieser, & Baldwin, 2010; Wallach, Murray, Salakhutdinov, & Mimno, 2009). The evaluation process focused on two primary components, which are the statistical model fit and the interpretability of the produced topics.

Likelihood measures and perplexity scores are two commonly used statistical measures assessing the appropriateness of the model. The likelihood measure identifies how the model fits the data (Griffths & Steyver, 2004). The likelihood measure could identify the most appropriate topic model with varying parameter settings (e.g., the number of topics) by comparing their average likelihood scores over the total number of iterations. The perplexity score evaluates the quality of the model by identifying how predictive the data is using the model. The perplexity score is computed as the inverse of the log-likelihood score, normalized by the number of words in the documents. Thus, the lower the perplexity score indicates that the data is highly predictable using the model.

In terms of the interpretability of the topic structure, topic models assign and provide a list of terms with their contributing weights to represent each topic. Using the vocabularies that are highly associated with the topic, one could attempt to understand and interpret the meaning of the topic. For instance, in the example topic assignment provided in Figure 6, the distinct meaning representation of each topic could be identified from their word distributions. One could expect topic 1 to represent main ideas regarding school and education, as it included keywords, such as “teacher”, “student”, and “classroom”. Such interpretability evaluation is often conducted manually but provides highly reliable evidence regarding the model performance. In this dissertation, both the statistical evaluation metrics as well as human evaluation will be adopted to identify the best topic model and validate its results.

Chapter Summary

Chapter 2 provided a general overview of the essential topics for my dissertation research. Previous AIG frameworks were categorized based on the use of item models or templates to generate test items. My literature review indicated that both the template-based AIG

returns highly usable and functional items in various domains. However, the complexity of designing appropriate cognitive models could be challenging in certain testing domains such as reading comprehension assessment. On the other hand, non-template-based AIG requires less human intervention to generate items from knowledge bases and source texts. But this approach depends on advanced statistical analysis techniques, such as natural language processing and deep neural network approaches, to generate test items. One of the commonly used techniques, topic modelling analysis using Latent Dirichlet Allocation, was also explained in detail. The next section described how the two AIG frameworks could be integrated to extend their capacities—this is this goal of my dissertation research.

Chapter 3: Methods

This chapter describes the dataset, model development frameworks, and system architecture that were used in my dissertation. To overcome the limitations of the previously introduced systems (Chapter 2), the current AIG framework extends the capacity of the template-based AIG with rich natural language processing techniques introduced in non-template-based AIG frameworks. Hence, in Chapter 3, I describe the components which were used to construct, implement, and demonstrate the current AIG framework in a previously challenging domain, reading inference-type item generation.

Data

The Harry Potter book chapters were used as the training documents to demonstrate my approach to generating inference-type items. The dataset consisted of 199 chapters from in the seven Harry Potter books - “The Philosopher’s Stone”, “The Chamber of Secrets”, “The Prisoner of Azkaban”, “The Goblet of Fire”, “The Order of the Phoenix”, “The Half-Blooded Prince”, and “The Deathly Hallows” (Table 2). This dataset was selected for two reasons. First, the stories are relatively well known, so that we could communicate the topic modelling results effectively. Second, the dataset included various semantic events and topics, which could make the topic modelling task more meaningful. Each chapter consisted of 2,121 to 12,022 tokens with 631 to 2,010 unique vocabularies. Figure 8 provides the distributions of the number of words, unique vocabulary, sentences in each chapter.

Interestingly, the number of tokens (i.e., length of the text) varied dramatically across the chapters while the amount of unique vocabulary in each chapter remained relatively similar across the chapters within each book. In terms of the frequent vocabularies, I identified that the word “the” was the most frequently occurring vocabulary across the chapters, appearing in a

chapter 260.32 times, on average. Other frequent words included functional words, such as “he”, “to”, “and”, and “she”. This finding indicates the importance of preprocessing the data by removing words that do not contribute to the meaning change of the context.

Table 2

A List of Example Book Chapters of the Current Dataset

Book 1 The Philosophers’ Stone	**Book 2** The Chamber of Secrets	**Book 3** The Prisoner of Azkaban	**Book 4** The Goblet of Fire
The boy who lived The vanishing glass The letters from no one … The man with two faces	The worst birthday Dobby’s warning The burrow … Dobby’s reward	Owl post Aunt Marge’s big mistake The knight bus … Owl post again	The riddle house The scar The invitation … The beginning

Book 5 The Order of the Phoenix	**Book 6** The Half-Blooded Prince	**Book 7** The Deathly Hallows
Dudley demented A Peck of owls The advance guard … The second war begins	The other minister Spinner’s End Will and Won’t … The white tomb	The dark lord ascending In memoriam The Dursleys departing … Epilogue

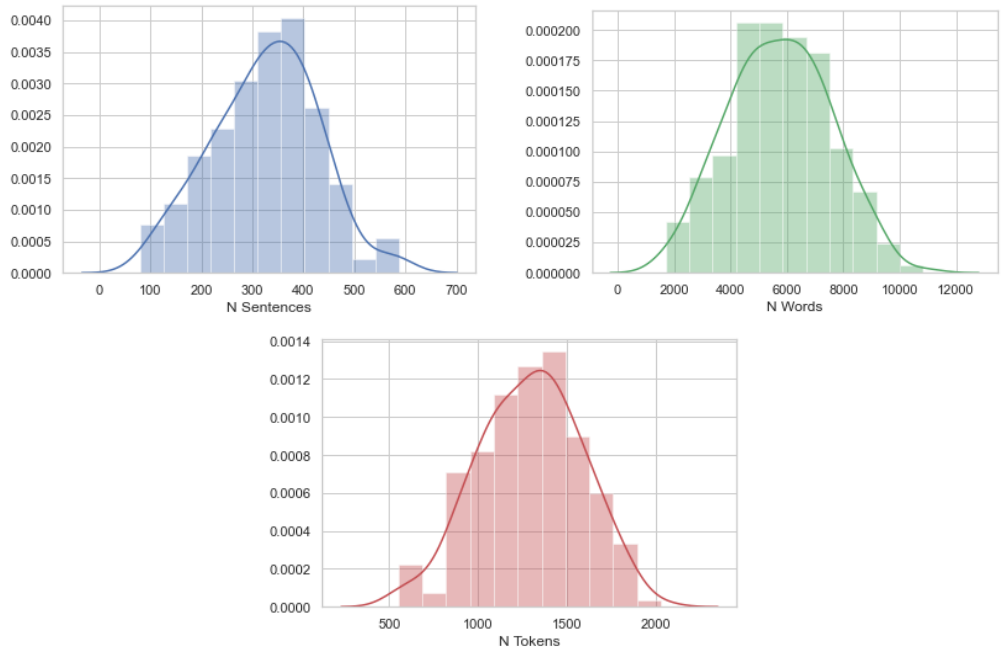


Figure 8. Distributions of the number of tokens and unique vocabularies in the chapters.

Analysis Framework Overview

Using the Harry Potter chapters as training documents of narrative stories, the item development analysis is conducted in three stages. Figure 9 presents a conceptual representation of the system framework. The first two stages investigate the ambiguous semantic structure of the text using advanced NLP techniques adopted from non-template-based AIG. The last stage apply item models from template-based AIG to generate test items from the disambiguated subtopic structures.

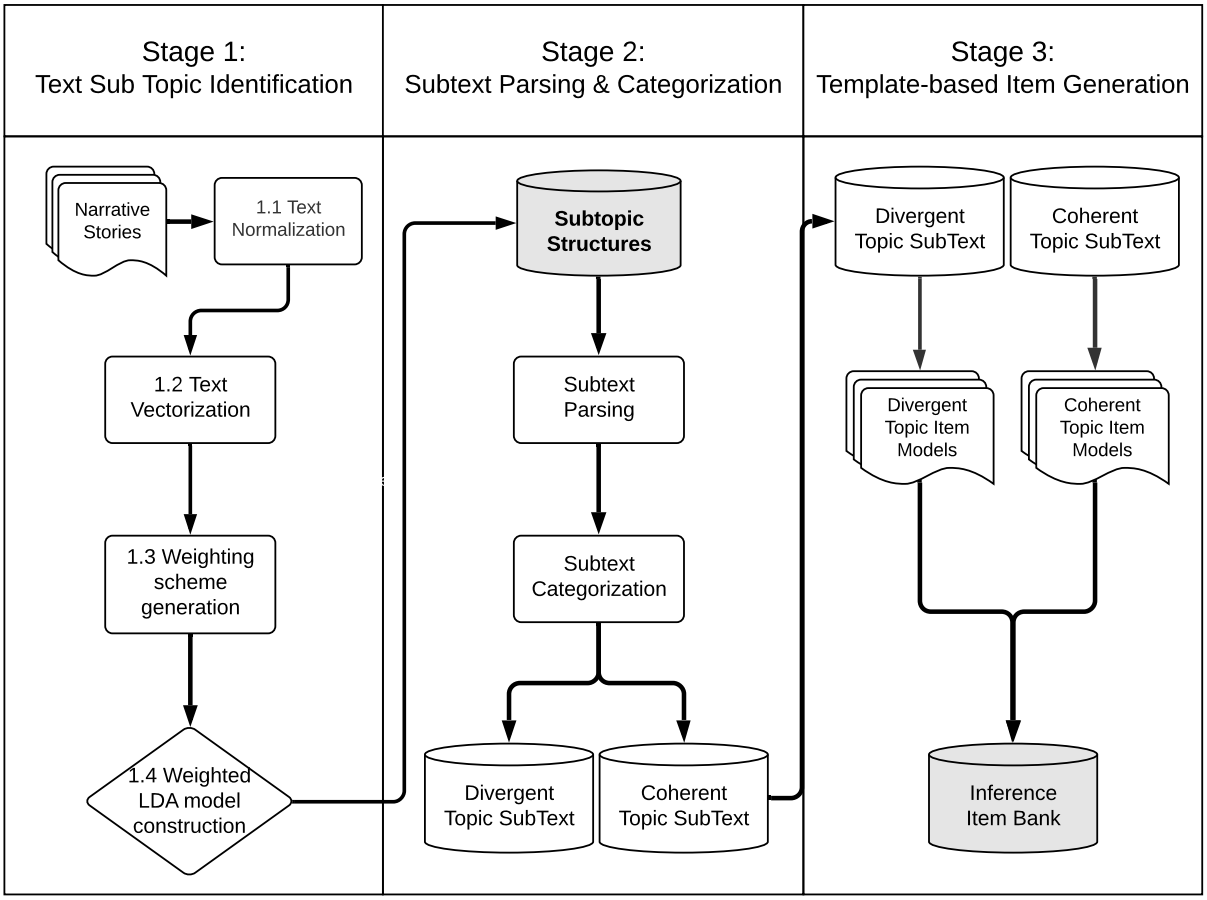


Figure 9. A conceptual overview of the system analysis framework.

The first stage of the analysis focuses on constructing topic models, which could effectively capture a comprehensive subtopic structure from the documents. A sentiment-

weighted latent Dirichlet allocation (LDA) algorithm is adopted which is used to identify the subtopic of the document. In this dissertation, the main topic of the text refers to the most evident and exterior structure of the text. This often corresponds to the main event or the chapter storyline. In contrast, subtopics refer to the secondary plots which could enrich the story with information about the characters' interactions and the sentiment. Hence, a weighted LDA model is constructed to capture the distinct topic structure of the main and the subtopics from the chapters.

Then, the second stage focuses on understanding and investigating the topic structure identified from the first stage and to categorize the reading passage candidates for item generation. The chapters will be parsed into smaller chunks of subtexts. Then, I will categorize the subtexts based on their topic structure similarity. This step is important to classify the reading passage candidates so that I could identify its comparability with the item-templates in the final stage.

The last stages involve item generation using the information about the subplot topics from Stage 1 and the information about the relationships between the topics from Stage 2. The last stage focuses on identifying and applying suitable item models to generate test items. The following sections introduce how each analysis is conducted with thorough technical details.

Stage 1: Sub-Topic Structure Identification using Weighted LDA

Stage 1.1: Data Preprocessing. A thorough text preprocessing will be conducted prior to topic modelling. This step is necessary to remove the noisy variations of words and phrases, thereby, improve the modelling accuracy to locate more distinguishable and interpretable topics in the next stages. Hence, this process focuses on normalizing the text by cleaning each token, or word, so that punctuations are considered as words. Then, the words are converted into original

forms (i.e. lemmas) and stem words using the NLTK package in Python 3 (Loper & Bird, 2002). This way, vocabularies sharing the same common or original forms will be grouped and analyzed together. Last, a list of English stop words from the NLTK package is used to identify and remove the redundant vocabulary from the text. Stopwords refer to a set of vocabularies that does not contribute to the significant meaning change of the text. For instance, a set of functional words – "the", "to", "of", "and" ¬ were identified as the most frequent words across the chapters in our current dataset. However, because the words do not contribute to the meaning change of the text, they are considered stopwords and removed from the texts in our analysis. A list of 179 English stopwords (e.g., "I", "me", "myself", "he", "the", "on", "of" etc.) provided by NLTK is adopted in the preprocessing stage.

Stage 1.2: Text Vectorization. Once the texts are preprocessed, they are transformed using the term-frequency inverse-document frequency (or TF-IDF) vectorization approach. In the conventional count-vectorization approach, texts are converted into a sparse numeric vector, which saves the count information of each word in the document. Hence, the word which appears frequently in the document will be represented with the highest value or the higher importance. Unlike the count-vectorization approach, the TF-IDF vectorizer accounts for the distinct words that appear in the documents. More specifically, TF-IDF vectorization provides the weights to the word count vector, by offsetting and downplaying the words and phrases that frequently appear across all the documents. The logic behind such a weighting scheme is that words that are too frequent across every document should not be contributing significantly to provide unique meaning structure to the text. The TF-IDF vectorization approach transforms the current dataset, the Harry Potter chapters, to construct document matrices. Then, the document matrices will be analyzed by the topic models in the next step. The resulting matrix of TF-IDF

vectorization will be represented with a row of documents, or chapters, and the columns of vocabularies presented across the chapters. The weighted values will represent the relative contribution of the specific word in the document to form a unique meaning structure.

Stage 1.3: Subtopic Modelling with the Weighted LDA Approach. One variational latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003) model will be introduced to identify the subtopic structures from the Harry Potter chapters. Unlike the previously introduced system which utilized the main ideas extracted from topic modelling approaches to generate higher-level reading items (e.g., Mazidi, 2017), I focused on identifying and modelling subtopics from the narrative stories to effectively generate inference items. As described earlier, subtopics in narrative stories focus on the interaction and the sentiment-related information in the text (Chatman, 1980; Murtagh, Ganz, & McKie, 2009). Thus, modelling subtopics from the documents could provide important evidence to generate items requiring examinees to "evaluate" and "assess" the overall outcome, sentiment, and interaction component of the story. This type of inference-type item was revealed to often assess the highest-level of inferential knowledge as reviewed in Chapter 2.

Modelling subtopics is not a trivial task. In narrative stories, subtopics that are related to the interaction and the sentimental components often lies below the main topic (Liu, Lv, Luo & Yang, 2009). This outcome occurs because the subtopics are often designed to enrich the story of the main event using the interactions between the characters and the sentiments of the main and the supporting characters. For instance, in Figure 10, we provided an example of topic-subtopic structures identified from Chapter 3, "The Letters from No One". It is important to notice that the subtopics often refer to the underlying interaction and the sentiment that are entailed as outcomes of the main event, or main topic.

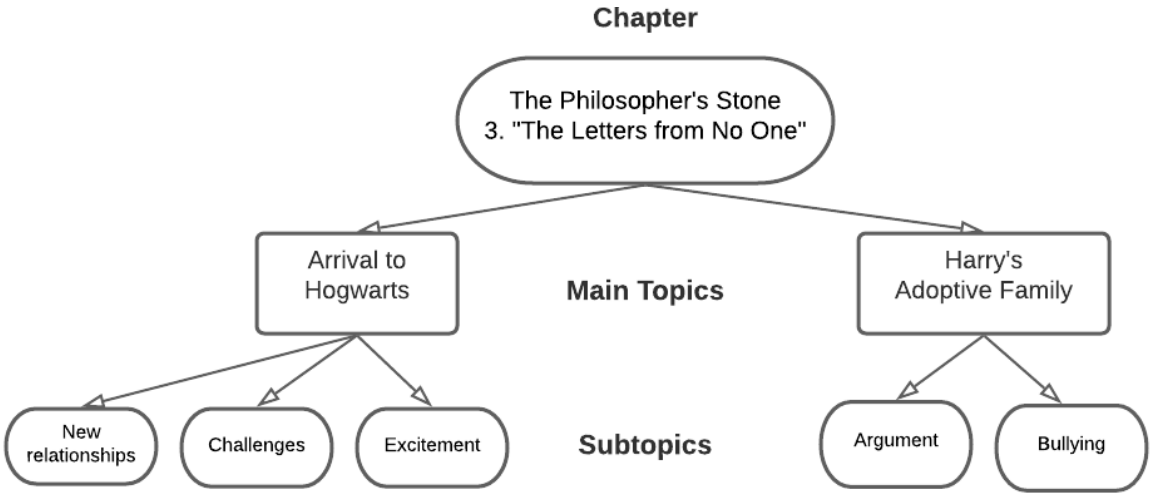


Figure 10. Example topic and subtopic structure in Chapter 3 “The Letters from No One”.

To effectively capture the subtopic structures from the current dataset, the proposed method adopts the second-level term weighting approach of Wilson and Chew (2010). The proposed weighting scheme allows the customized term weighting in collapsed Gibbs sampling to estimate the word-topic (ϕ) and the topic (θ) distributions. In the original algorithm, the number of tokens or the count information of the word is used to estimate the topic structures. Instead, the term-weighting scheme of Wilson and Chew (2010) allows the users to provide empirical weights (e.g., Point-wise mutual information) to the word-count information to estimate the important parameters defining the topic structures. With this approach, one can systematically replace the word-count matrix with a weighted-term matrix, M_{nmk}. This weighted term matrix, M_{nmk}, will be designed based on each vocabulary contribution to the overall sentiment and the interaction of the event. The term-weighting matrices and their development will be described in the next sections.

Stage 1.4: Sentiment Score Weighted LDA for the Subplot Modelling. In this study, the subtopics refer to the underlying stream of stories that focus on the characters’ interactions and sentiment. To provide a comprehensive weighting scheme that represents the definition of

subtopic structures, two natural language processing techniques will be introduced, which are named-entity extraction and sentiment analysis. First, named-entity extraction is conducted to identify and extract the parts of the text that indicates interactions between the characters. This can be achieved by locating the sentences in the document which included more than one named entity names of the characters in this study. I hypothesized that if a sentence includes more than two characters' names, it indicates that the sentence presents information about the relationship or the interaction happening between the characters. Once the parts of the text that identify "interactions" components are extracted, we will focus on identifying the vocabulary-weights that represent the sentimental contribution of each word in the extracted documents.

The Pearson correlation between the word count information of each sentence in the extracted parts of the document is computed, with their corresponding sentence-level sentiment scores. In essence, the correlation coefficient could determine the relationship between the frequency of a specific word in the sentence and the compound sentiment score of the sentence. The sentence-level sentiment score will be estimated using a lexical, rule-based sentiment analysis system, the Valence Aware Dictionary and Sentiment Reasoner, or VADER (Hutto & Gilbert, 2014). VADER uses a list of sentiment lexicon with the information about the sentiment directionality of each lexicon. Using the set of pre-defined rules with the lexical directionality information, the system provides a continuous score range from -1 to 1 to identify the compound sentiment score. In general, VADER compound score greater than 0.05 indicates a strong positive sentiment of the text, while a score less than -0.05 indicates a strong negative sentiment of the text (Hutto & Gilbert, 2014). Next, the absolute values of the correlation coefficient are retrieved. The last step inspects whether any function words or vocabularies were assigned high weights due to their high occurrence across all documents. This will be controlled by acquiring a

list of neutral words based on the VADER sentiment score and suppressing the weights of the corresponding words to zeros. This way, term-weights of the high-frequency word such as “the”, “he”, and “she”, will be suppressed.

This weighting scheme will provide an alternative way of suppressing the contribution of the high-frequency words with little sentimental contribution to the text. Thus, the estimation of the topic distributions of θ and ϕ and the Gibbs sampling with the updated sentiment-focused term-weighting scheme, M, by replacing the count vectors as in Equation 8.

$$P\left(\mathrm{z_{n,m}} = \mathrm{k} \middle| \mathrm{z_{-(n,m)}}, \alpha, \beta\right) \propto \frac{\left(M^{-(\mathrm{n,m})}_{(\mathrm{n,\cdot,k})} + \beta\right)}{M^{-(\mathrm{n,m})}_{(\cdot,\cdot,\mathrm{k})} + \mathrm{V}\beta} \cdot \frac{\left(M^{-(\mathrm{n,m})}_{(\cdot,\mathrm{m,k})} + \alpha\right)}{M^{-(\mathrm{n,m})}_{(\cdot,\mathrm{m},\cdot)} + \mathrm{K}\alpha}, \text{ where} \quad (8)$$

K: the total number of topics

V: the total number of vocabularies

α: the Dirichlet parameter setting the topic distribution for the documents

β: the Dirichlet parameter setting the topic distribution for the words.

Stage 2: Subtext Parsing and Categorization

Once the subtopic structures are identified from the chapters in Stage 1, Stage 2 focuses on parsing the chapters into subtexts of shorter lengths. This is necessary to locate the reading passage candidates that are suitable for item generation. More specifically, the chapters will be parsed into subtexts of close to 400 to 500 words to adhere to the current high-stakes large-scale reading assessment item format (Figure 3). Then, the two categories of subtexts will be identified based on their subtopic mixture distributions: “coherent-topic” texts “divergent-topic” texts. This categorization is necessary to provide suitable item models subtexts with different subtopic mixture distributions in the next stage.

The parsing and the categorization are conducted using a rule-based system as presented in Figure 11. To iteratively parse the subtexts and investigate their topic structure, a sliding window is identified, which could contain a list of sentences with approximately 500 words. As the sliding window moves across a total of 199 chapters, it will compute the total topic-weight scores of the subtext. This will be identified using the document-topic matrix identified in Stage 1. The combined topic-weights will be used to filter the subtexts with no prominent topic structure. If the total topic-weight of the subtext is lower than the cut-off score, then we will remove the subtext from the candidate pool. Once the subtexts with prominent topic structures are located, we will evaluate whether the topics presented within the texts are coherent or divergent. “Coherent-topic” subtexts indicate that the introduced topics are of high similarity, conveying homogenous sentimental subtopics. Conversely, “divergent-topic” subtexts indicate distinct sentimental components, and ideas are provided as subtopics in the document. To classify the texts into these categories, we will first investigate the number of dominant topics in the subtext.

For instance, if one of the topic in the text composed of more than 70% of the overall topic structure, then the text was classified as coherent topic category. If the subtext presented more than one dominant topic, then the similarity of the topics presented in the subtexts were evaluated using the Jennsen-Shannon divergence measure (Fuglede & Topsoe, 2004). Jensen-Shannon divergence measure is a symmetric and smoothed version of Kullback-Leibler divergence, which compares the similarity of the difference between two or more probabilitis (Equations 9 and 10).

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)}\right) \tag{9}$$

$$\mathrm{JSD}\,(\mathrm{P} \parallel \mathrm{Q}) = \frac{1}{2}\,\mathrm{D}\,(\mathrm{P} \parallel \mathrm{M}) + \frac{1}{2}\,\mathrm{D}\,(\mathrm{Q} \parallel \mathrm{M}),\ \text{where}\ \mathrm{M} = \frac{1}{2}(\mathrm{P} + \mathrm{Q}). \tag{10}$$

The Jensen-Shannon divergence could range from 0 to 1, in which 0 indicates that the two distributions are equal and 1 indicating that the two distributions are distinct. Hence, the Jensen-Shannon divergence was used to investigate by how highly the topics are related to each other based on their topic-term distributions. Divergence measure close to 1 would mean that the presented topics will indicate the text contains "divergent" topics.

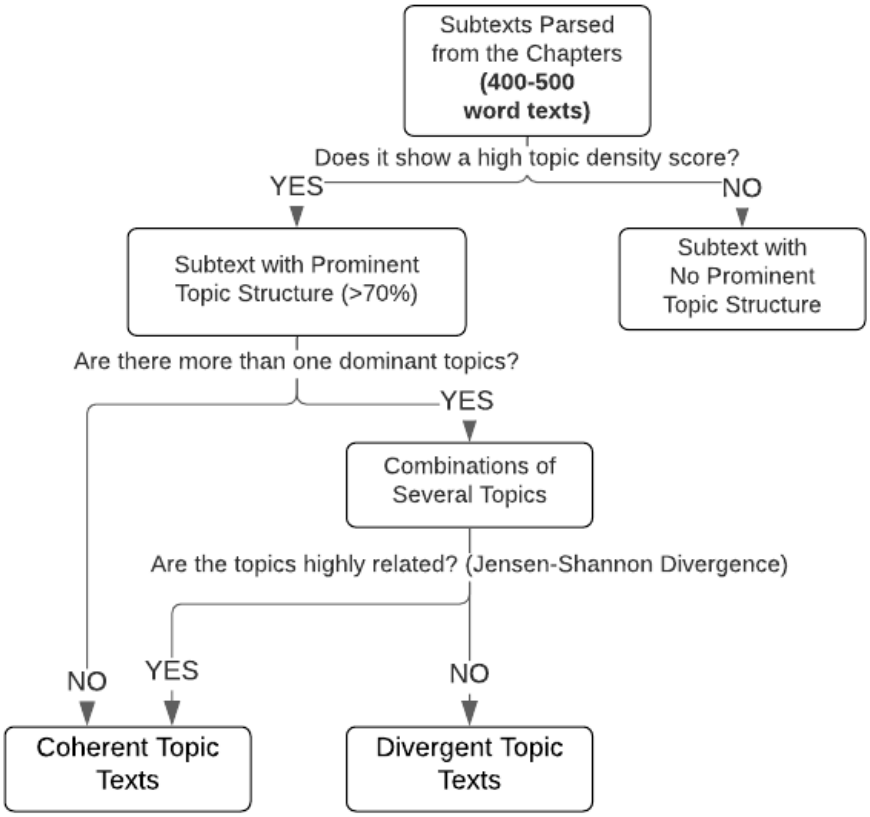


Figure 11. A conceptual representation of a rule-based subtext categorization.

To summarize, Stage 2 presented a rule-based model that could parse the chapters into 500-word length subtexts and classify them based on their topic structures – texts with coherent topics and divergent topics (Figure 11). The categorized subtexts will be provided with varying types of appropriate item models in Stage 3 to generate reading inference items.

Stage 3: Reading Inference-type Item Generation

The proposed system uses the template-based item generation approach in Stage 3. Template-based AIG uses item models, which predefine a set of components that could be applied universally with various conditions to generate items (Gierl & Lai, 2013). Item models are customized for the candidate subtexts of varying topic structures. More specifically, four types of item models will be constructed and applied to generate inference items.

Stage 3.1: Item Model generation. Four item models for coherent- and divergent-topic passages were created. Each item model featured four components, which are the incomplete question-statement (or stem), distractor-selection mechanism, answer selection mechanism, and the appropriate types of reading passages. Figures 12, 13, 14, and 15 present four item models. Each item model was developed to produce test items given appropriate categories of reading passages: coherent-topic and divergent-topic texts.

Item models logically combined the featured components to assess the examinees' ability to make correct inferences from the given text. Each item model assess whether "the examineess could correctly identify the parts of the text (sentences) coherent with the given sentimental keyword" (coherent item model 1), "the examineess could identify the sentiment-topic keywords based on the highlighted parts of the text (sentences)" (coherent item model 2), "the examineess could distinguish the varying sentiment between the different parts of the text and represent them as topic keywords" (divergent item model 1), and "the examineess could indicate parts of the text presenting the different sentimental topic" (divergent item model 2).

For instance, the template in Figures 12 and 13 takes in coherent-topic texts. In Figure 12, if the reading passage has shown the highest association with Topic A, then the rest of the components, such as the answer keys and the distractors were located from the component of

Topic A: the Topic A key words and Topic A key sentences. Given that Topic A included a list of keywords –"lucky", "excited", "generous", and "hopeful"–, then this could generate four sets of the stem sentence: "The main character's feeling "lucky" is most likely related to the statement", "The main character's feeling "excited" is most likely related to the statement", "The main character's feeling "generous" is most likely related to the statement", and "The main character's feeling "hopeful" is most likely related to the statement". Followed by the stem, the answer key could be identified by locating the sentences in the text with high overall topic weights (or topic key sentences). Conversely, the distractors could be extracted by locating the sentence that shows low topic weights.

Figures 14 and 15, on the other hand, take in divergent-topic texts. In Figure 14, the appropriate input reading passage has shown the highest associations with two distinct topic, Topic A and B. Thus, the key elements of Topic A and B (i.e., topic key sentences and keywords) are used to construct the item elements, stem and a correct option. The incorrect options are, then, identified from the rest of the topic structures except Topic A and B (or A^{-}, B^{-}).

In summary, the coherent topic item models intended to evaluate examinees' ability to make correct inferences by associating relevant sources of evidence reflecting the overall "sentiment" given the text (Figures 12 and 13). On the other hand, the divergent topic item models intended to evaluate examinees' ability to discern textual evidence representing unassociated sentimental topics (Figures 14 and 15). The two major textual evidence–topic keywords and topic key sentence–act as important item generation components.

Coherent Topic Item Model 1

Input:
[Coherent topic text] [Topic A Sentences] [Topic A Keyword] [Topic A^- Sentences] [Characters]

Reading passage:
[Coherent topic text]

Stem:
The main character's [Characters] feeling [Topic A Keyword] is most likely related to the statement:

Options:
Answer: [Topic A Sentences]
Distractors: [Topic A^- Sentences]

Figure 12. Example Item Model for coherent-topic texts (1).

Coherent Topic Item Model 2

Input:
[Coherent toic text] [Topic A Sentences] [Topic A Keyword] [Topic A^- Sentences] [Characters]

Reading passage:
[Coherent topic text]

Stem:
What can be reasonably inferred from [Topic A Sentences] that the main character [Characters] felt:

Options:
Answer: [Topic A Keyword]
Distractors: [Topic A^- Keywords]

Figure 13. Example Item Model for coherent-topic texts (2).

Coherent Topic Item Model 2

Input:
[Divergent toic text] [Topic A Sentences] [Topic B Sentences]
[Topic A Keyword] [Topic B Keyword] [Topic AB⁻ Keywords]

Reading passage:
[Divergent topic text]

Stem:
How is the main character's sentiment described in [Topic A sentence] different from [Topic B sentences]?

Options:
Answer: [Topic A keywords] and [Topic B keywords]
Distractor: [Topic A keywords] and [Topic AB⁻ keywords]
Distractor: [Topic A keywords] and [Topic AB⁻ keywords]
Distractor: [Topic AB⁻ keywords] and [Topic B keywords]
Distractor: [Topic AB⁻ keywords] and [Topic B keywords]

Figure 14. Example Item Model for divergent-topic texts (1).

Divergent Topic Item Model 2

Input:
[Divergent toic text] [Topic A Sentences] [Topic B Sentences] [Topic C Sentences]
…

Reading passage:
[Divergent topic text]

Stem:
Which of the following statements indicates different sentiment from the others?

Options:
Answer: [Topic A Sentences] & Distractor: [Topic A⁻ Sentences]
Answer: [Topic B Sentences] & Distractor: [Topic B⁻ Sentences]
Answer: [Topic C Sentences] & Distractor: [Topic C⁻ Sentences]

Figure 15. Example Item Model for divergent-topic texts (2).

Chapter Summary

Chapter 3 described the current AIG framework that was adopted in my dissertation. The AIG framework focuses on uncovering underlying subtopic structures of texts to generate new test items. Various analyses based on natural language processing, such as a topic modelling approach , was introduced to identify critical text features needed to locate subtopics. It is noted that a modified latent Dirichlet allocation analysis could effectively identify the sublayers of topic structures in narrative stories. Also, the system introduced a rule-based approach to categorize texts based on their subtopic models. Last, item models from the template-based approach were developed to generate items by assembling the information acquired from the topic modelling stage.

Chapter 4: Results

This chapter focuses on communicating and explaining the results of the three primary stages in my AIG system: Stage 1: sentiment-weighted topic modelling, Stage 2: subtext candidate parsing, and Stage 3: test item generation with item models. Stage 1 results identify the characteristics and the quality of sentiment-focused topic modelling results by comparing the topic keywords with naïve topic modelling system. Stage 2 describes the rule-based parsing results in proportioning the candidate subtexts based on their topic mixture distributions. Stage 3 identifies the specific examples of test items generated from the four item models with the subtext candidates validated from Stage 2. This chapter concludes by providing results for a customized validation analysis of each stage in the item generation process. The validation analyses along with their findings provide evaluation evidence on the important stages of the AIG framework to assure the generated items are of high quality.

Result of Stage 1: Unweighted and Sentiment-Weighted LDA Topic Model Results

The final weighted LDA model included a total of ten topics. The grid search of the best parameter of the LDA model revealed that with the learning decay of 0.70 and the number of topics of 10, the best model could be located with the perplexity of 93.95. Table 3 and Figure 16 provide a comparison of the LDA models with varying parameter settings based on their log-likelihood score. The final ten topics generated from the unweighted and sentiment-weighted LDA topic models produced interpretable results. Each topic produced keywords with a clear and coherent interpretation. The topic keywords with the highest contributing weights were presented and interpreted in an attempt to label the extracted topics in the unweighted and the sentiment-weighted LDA topic models, as described in the next sections.

Table 3

Log-likelihood based on the number of topics and learning decay

Learning Decay	Number of Topics						
	2	**5**	**10**	**15**	**20**	**25**	**30**
0.5	-97084.86	-96716.18	-96099.10	-96346.67	-96716.35	-96707.17	-97365.50
0.7	-96763.89	-96348.40	-95899.44	-96474.47	-96749.08	-96831.12	-96958.85
0.9	-96922.16	-96527.48	-96393.59	-96616.41	-96616.40	-97190.29	-97473.65

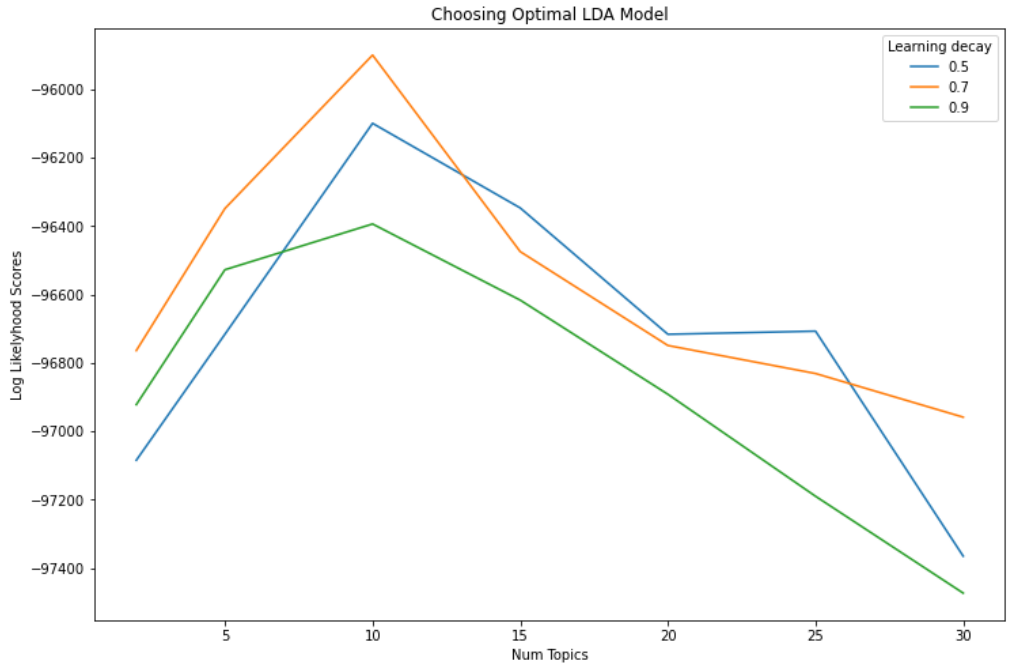


Figure 16. Optimal LDA Model and the Parameter settings.

Unweighted LDA Topic Model Results. Table 4 provides keywords of the ten topics extracted from the unweighted LDA topic model. Topic 4 with the keywords such as “Team”, “Snitch”, “Broom”, “Firebolt”, “Field”, and “Buckbeak” can be labelled as “Quidditch”. Topic 6 with the keywords such as “Fudge”, “Minister”, “Dementors”, and “Azkaban” can be labelled as “Ministers”. Topic 7 included keywords such as “Kreacher”, “Bellatrix”, “Master”, and “Workmtail”, thus, can be labelled as “Horcrux”. The combinations of extracted topics were then

located in various parts of the Harry Potter texts. This step was conducted by computing the posterior distribution of the topic distribution over each document in the Harry Potter chapters.

Table 4

Unweighted LDA Topic Model Results

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Lockhart	Kitchen	Moody	Yeh	Umbridge
Flich	Tonks	Crouch	Ter	Trelawney
Luna	Prophet	Cedric	Wood	Cho
Goyle	Mundungus	Krum	Team	Lesson
Crabbe	Hedwig	Bagman	Snitch	Parvati
Peeves	Christmas	Diggory	Broom	James
Nick	Moody	Fluer	Firebolt	Homework
Seamus	Albust	Tournament	Field	Angelina
Headless	Arthur	Madame	Yer	Lavender
Ravenclaw	Bedroom	Maxime	Buckbeak	Divination

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Fudge	Kreacher	Slughorn	Dobby	Uncle Vernon
Minister	Bellatrix	Riddle	Elf	Dudley
Dementors	Master	Sir	Master	Aunt
Scabbers	Workmtail	Bathroom	Elves	Petunia
Crookshanks	Sword	Quirrell	Squeaked	Dursleys
Shop	Eater	Map	Bludger	Kitchen
Azkaban	Snake	Lesson	Tea	Car
Trunk	Luna	Lavender	Clothes	Drive
Cornelius	Horcrux	Soul	Kitchen	Privet

Sentiment-weighted LDA Topic Model Results. Compared to the unweighted LDA topic modelling results, the sentiment weighted-LDA outcomes were represented with the top keywords or the set of vocabulary related to the main characters' sentiment and interaction with the high contribution to explain the topic (Tables 5, 6, and 7). For instance, Topic 1 included keywords representing positive interactions and sentiments between the characters related to excitement, such as "Excellent", "Bright", "Excitement", "Hopeful", and "Wonderful".

Similarly, Topic 5 presented keywords that are related to the emotions of caring, such as “Love”, “Powerful”, “Truly”, “Kindly”, “Favourite”, and “Loved”. Topic 3, on the other hand, presented keywords showing negative sentiment keywords, such as “Terrible”, “Hatred”, “Attack”, “Horrible”, and “Miserably”. Topic 6 included keywords, which indicate sentiment keywords related to anger with the keywords, such as “Fury”, “Anger”, “Rage”, and “Destroy”. The rest of the topic categories consistently provided sentiment-related keywords that provide coherent interpretation. The keywords showed dramatic changes compared to the unweighted LDA results, which mostly focused on the main events, plots of the story, specific words and terminology, including the names of the main characters associated with the events. These outcomes were expected, as the main event-related vocabulary tends to occur more frequently compared to the others to describe the main topic and located using the unweighted LDA topic results. Moreover, providing a sentiment-weighted scheme to compute the topic distributions tend to help address the issues of highly frequent words by highlighting the contribution of interaction and sentiment related terminology in topic-word and document-topic distributions in the LDA topic model computations (Table 6 and Table 7).

Table 5

Sentiment-weighted LDA Topic Model Results

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Excellent	Party	Terrible	Attacked	Love
Bright	Saved	Hatred	Attack	Powerful
Definitely	Love	Attack	Impatiently	Truly
Interested	Lucky	Fault	Badly	Kindly
Excitement	Hopeful	Rage	Panic	Favourite
Hoping	Cheerful	Fault	Trapped	Loved
Wonderful	Enjoying	Horror	Challenging	Cared
Strong	Hopefully	Awful	Poor	Clever
Excitedly	Interested	Horrible	Horrified	Surely
Applause	Delighted	Miserably	Ashamed	Touched
Enjoying	Perfectly	Problematic	Crying	Genuine
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Fury	Die	Terrified	Welcome	Charms
Anger	Injured	Worried	Pleasant	Magical
Rage	Ill	Scared	Delighted	Lucky
Destroy	Killing	Frightened	Amazed	Cheering
Fake	Murdered	Nasty	Brave	Winning
Forbidden	Accident	Dangerous	Relief	Accomplish
Stolen	Dangerous	Guilty	Proud	Divination
Irritated	Struggling	Trapped	Favorite	Victory
Terrified	Drowning	Danger	Cheers	Achievement
Scream	Sobbing	Fooled	Supportive	Luck
Furious	Terrified	Weak	Friendly	Delighted

Table 6

Parsed document-topic distribution table of the sentiment-weighted LDA results

	Topic										Dominant Topic
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Chapter 1	0.00	0.00	0.00	0.33	0.00	0.13	0.19	0.30	0.00	0.05	Topic 4
Chapter 2	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.27	0.00	Topic 3
Chapter 3	0.00	0.00	0.00	0.34	0.00	0.04	0.07	0.00	0.46	0.08	Topic 9
Chapter 4	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.18	0.61	0.00	Topic 9
Chapter 5	0.00	0.00	0.22	0.00	0.00	0.12	0.03	0.21	0.43	0.00	Topic 9
Chapter 6	0.34	0.00	0.00	0.20	0.09	0.06	0.00	0.31	0.00	0.00	Topic 1
Chapter 7	0.13	0.00	0.14	0.00	0.00	0.00	0.04	0.00	0.69	0.00	Topic 9
Chapter 8	0.13	0.00	0.76	0.00	0.00	0.00	0.00	0.05	0.00	0.05	Topic 3
Chapter 9	0.00	0.00	0.08	0.00	0.00	0.00	0.14	0.29	0.06	0.43	Topic 10
Chapter 10	0.16	0.00	0.13	0.00	0.00	0.27	0.00	0.18	0.00	0.26	Topic 6
Chapter 11	0.24	0.00	0.06	0.02	0.00	0.00	0.00	0.12	0.00	0.55	Topic 10
Chapter 12	0.31	0.08	0.00	0.40	0.11	0.00	0.00	0.00	0.05	0.05	Topic 4
…						…					

Table 7

Parsed topic-word distribution table of the sentiment-weighted LDA results

	Topic									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Excellent	0.75	0.13	0.00	0.00	0.00	0.00	0.03	0.00	0.11	0.11
Bright	0.65	0.02	0.04	0.00	0.03	0.00	0.14	0.02	0.02	0.07
Definitely	0.79	0.01	0.00	0.07	0.06	0.00	0.00	0.02	0.00	0.04
Interested	0.66	0.06	0.00	0.00	0.00	0.13	0.06	0.09	0.00	0.00
Excitement	0.67	0.02	0.00	0.00	0.06	0.11	0.03	0.00	0.03	0.09
Hoping	0.49	0.13	0.14	0.10	0.02	0.00	0.05	0.00	0.00	0.08
Wonderful	0.76	0.00	0.06	0.06	0.06	0.00	0.00	0.00	0.06	0.00
Pretty	0.49	0.01	0.09	0.09	0.12	0.18	0.00	0.03	0.00	0.00
Strong	0.64	0.00	0.07	0.08	0.15	0.05	0.00	0.00	0.00	0.00
Excitedly	0.70	0.03	0.03	0.16	0.00	0.00	0.00	0.00	0.03	0.05
…						…				

Specific topic distributions for the parts of the Harry Potter texts could also be evaluated. Tables 8 and 9 provide example topic distribution of Book 3 – Chapter 20, “The Dementor's Kiss” and Book 1 – Chapter 11 “Quidditch”. The results indicated that the sentiment-weighted LDA topic modelling could successfully extract the topic keywords that are highly related to the sentimental aspects of the text. For instance, “The Dementor”s Kiss” is a chapter which focuses on the story evolving around the act of a “Dementor”, which could take away one’s soul. Hence, many of the interactions among the characters in this chapter present negative and horrifying sentiment. Such aspects of the subtopics were extracted and captured as Topic 3, 4, 6, and 7 (Table 8). As presented in Table 8, these topics all attempt to communicate consistent sentiment that underlies the character’s interaction. Conversely, the chapter “Quidditch” included both positive (Topic 10) and negative (Topic 4) sentimental aspects, which represent the nature of the chapter focusing on the competitiveness and friendship, losing and winning of the characters in Quidditch match (Table 9).

Table 8

Sentiment-weighted LDA Topic Results of “Book 3-Chp.20: The Dementor’s Kiss”

Topic									
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.01	0.01	0.20	0.25	0.01	0.22	0.27	0.01	0.01	0.01

- **Topic 3**: “Terrible”, “Hatred”, “Attack”, “Fault”, “Awful”, “Horrible”, “Miserable”
- **Topic 4**: “Attacked”, “Attack”, “Impatiently”, “Badly”, “Panic”
- **Topic 6**: “Fury”, “Anger”, “Rage”, “Destory”, “Forbidden”, “Stolen”, “Painful”
- **Topic 7**: “Die”, “Ill”, “Injured”, “Murdered”, “Killing”, “Accident”, “Dangerous”

Table 9

Sentiment-weighted LDA Topic Results of "Book 1-Chp.11: Quidditch"

Topic									
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.01	0.01	0.01	0.40	0.01	0.01	0.01	0.01	0.01	0.52

- **Topic 4**: "Attacked", "Attack", "Impatiently", "Badly", "Panic"
- **Topic 10**: "Charms", "Magical", "Lucky", "Cheering"

Results of Stage 2: Subtext Parsing Results

Recall a subtext of 400 to 500 words was used to adhere to the current high-stakes large-scale reading assessment item format (Chapter 3). A total of 31,061 subtexts of 400 to 500 words were generated using the sliding window approach. An average of 183 candidate subtexts was generated using the sliding window with a size of 400 to 500 words from each chapter. Among the subtext candidates, a total of 6,065 sample texts were removed as they did not achieve the topic distribution density score above zero. The topic distribution density score was computed by locating the key sentences in a subtext which are composed of topic keywords with the corresponding topic weights. Thus, density score of zero indicates the subtexts did not include any vocabulary that was identified as topic keywords in our sentiment-focused LDA results. Conversely, the text with the key sentences consisting of topic keywords of diverse topic structures would show high topic density score.

Removing subtexts with no topic key sentences resulted in a total of 30,347 candidate subtexts, which were further categorized based on the contextual similarities between the topic mixtures. Figure 17 provides how the topic mixture distributions varied in the example candidate subtexts computed as the sliding window applied to the different parts of the Harry Potter

chapters. Figure 18 provides the total number of candidate subtexts based on their most dominant topic.

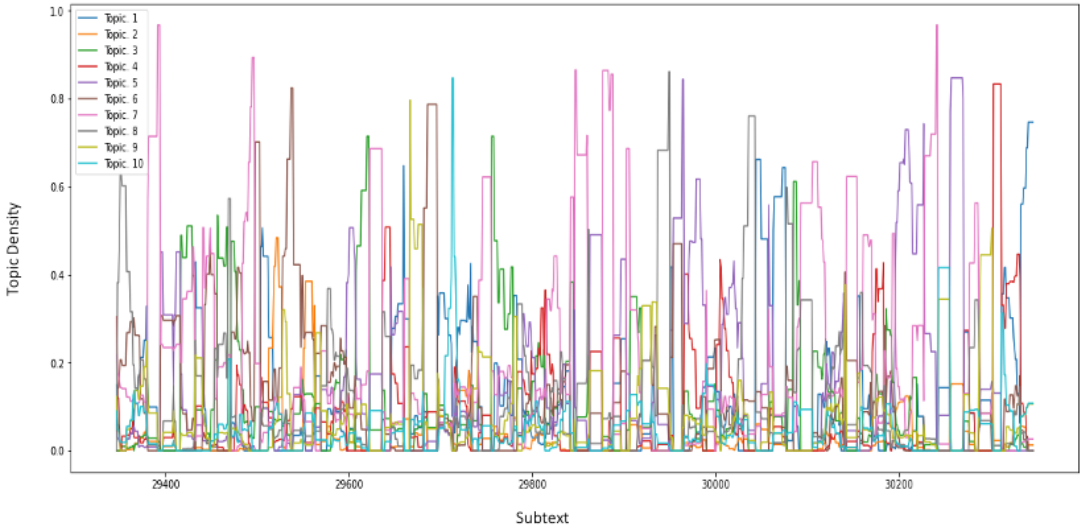


Figure 17. Topic mixture distributions of the sample parsed texts to serve as reading passages.

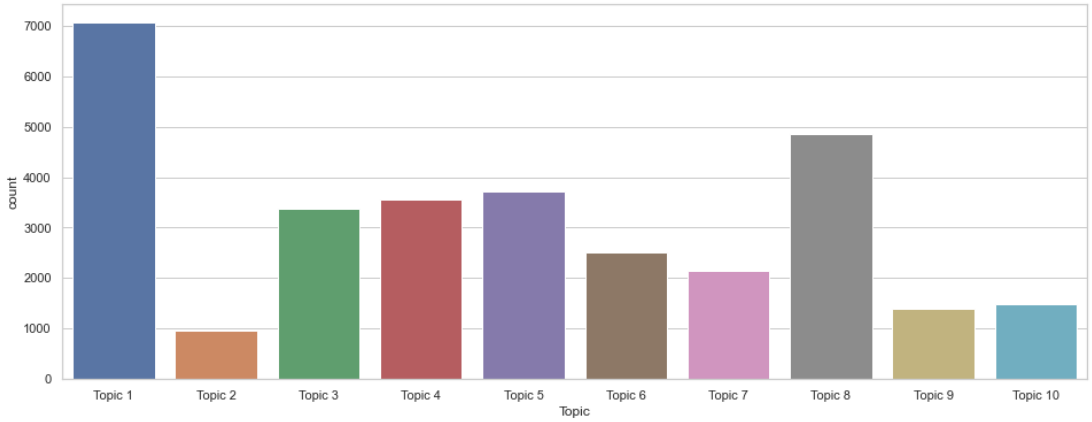


Figure 18. Total number of candidate subtexts with their dominant topic categories.

The subtexts with one dominant topic structures were identified and categorized as coherent texts. The candidate texts with one dominant topic explained more than 70% of the total topic mixture, as I defined in the rule-based parsing model in Chapter 3 (p. 41). This resulted in a total of 3,408 subtexts classified as one-dominant topic structures, thus, providing a coherent topic structure in the text. For instance, Figure 19 provides topic distributions of the example subtexts with one dominant topic structure (yellow) and the divergent topic structure (green). As

the distribution indicates, the coherent topic text included at least one topic dominating more than 70% of the overall topic structure (Topic 10, yellow). On the other hand, divergent topic texts present more than two topics explaining the overarching story of the text with a fair amount of topic contributions.

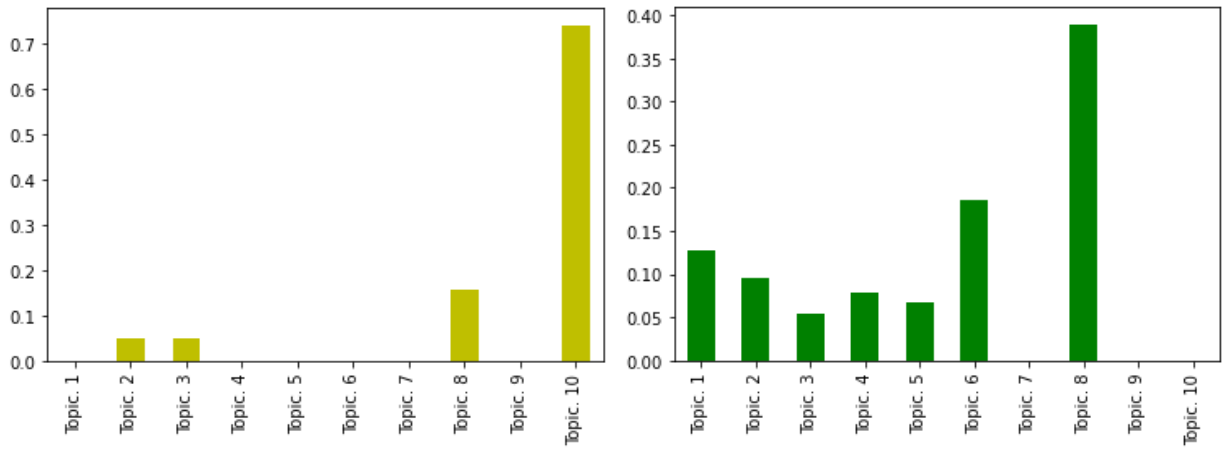


Figure 19. Topic distributions of the example subtexts from coherent and divergent categories.

Next, a Jessen-Shannon divergence was computed to measure the similarities between the topics based on their word distributions. The results indicated that none of the generated topics featured highly similar context. Table 10 provides the Jessen-Shannon divergence measures comparing the word-distribution among the ten topics. The measure identifies similar probability distributions by providing a value closer to 0 when the two distributions are equal. A value closer to 1, in turn, indicates that the two distributions are distinct. The outcome from this analysis produced the remaining candidate texts being categorized as divergent topic texts (N= 26,939). This resulted in a total of 8% of the candidate subtexts being classified as coherent texts with one dominant topic (N = 2,604) and the rest of the candidate texts as divergent texts with several distinct topic mixtures (N= 28,457). Figure 20 provides an overview of the framework applied to parsed the candidate subtexts in the current dataset.

Table 10

Jessen-Shannon Divergence Measure of Topic-Word distributions

	Topics									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1										
2	0.86									
3	0.82	0.89								
4	0.80	0.87	0.82							
5	0.85	0.86	0.85	0.84						
6	0.82	0.92	0.82	0.85	0.81					
7	0.87	0.91	0.81	0.89	0.85	0.84				
8	0.87	0.89	0.87	0.88	0.83	0.84	0.84			
9	0.78	0.92	0.81	0.82	0.81	0.83	0.86	0.84		
10	0.82	0.82	0.84	0.86	0.88	0.86	0.86	0.84	0.84	

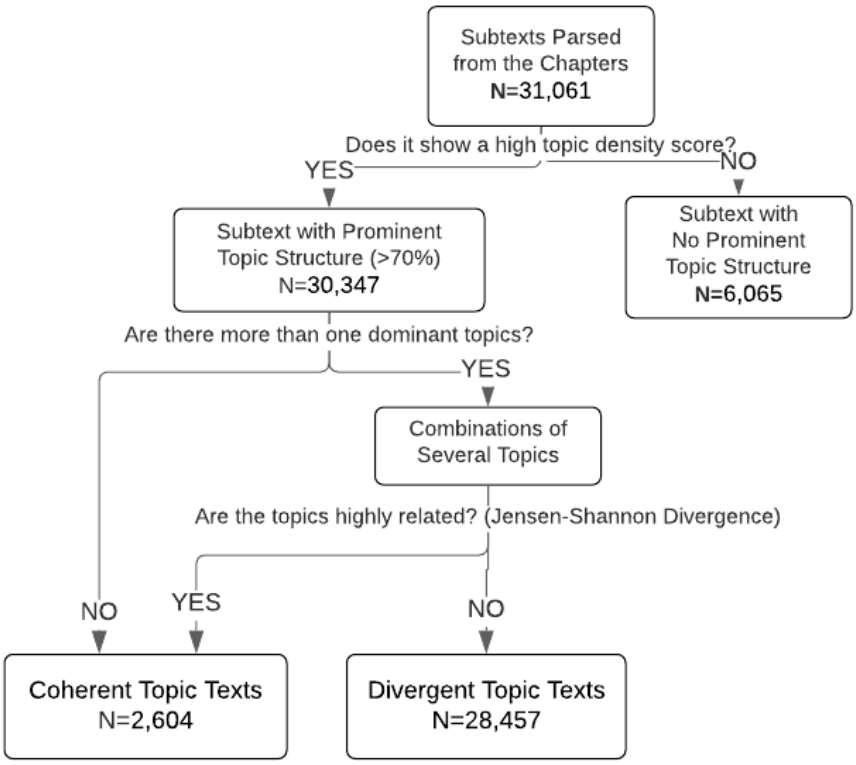


Figure 20. Final topic categorizaition results based on the rule-based model.

Results of Stage 3: Item Model Application Results

Four item models were generated and applied to the subtexts of divergent and coherent categories. Table 16 provides the descriptive statistics of test item variations which were generated given a candidate reading passage or a subtext candidate. On average, coherent item models could generate more variational keyed options and distractors compared to the divergent item models. The second coherent item model could generate an average of 3.22 keyed option and 8.45 distractors given a source text or a reading passage. Note, however, that the first divergent item model could only generate an average of 0.72 distractors per reading passage. This resulted in a surprisingly smaller number of generated items (N= 1,071) with the adequate number of the keyed option (≥ 1) and distractors (≥ 3) based on the conventional MC test items in operational administration standards (ACT, 2020; College Board, 2020; Downing, & Haladyna, 1997).

The divergent item models, on the other hand, showed a relatively lower average for the variational keyed-option and distractors that were generated per reading passage (Table 11). However, most of the generated items included an adequate number of the keyed option (≥ 1) and distractors (≥ 3) based on the conventional MC test items. This resulted in a fairly large number of test items generated from the divergent item model 1 (N= 28,457) and model 2 (N= 29,206) given the total divergent topic candidate subtexts. In the next sections, I provided more specific information about how the items were constructed using coherent and divergent item models using real examples.

Table 11

Item Generation Statistics per Item Model

	Coherent Item Models		Divergent Item Models	
	1	**2**	**1**	**2**
Average number of keyed option	3.35	3.22	1.00	2.92
(STDEV)	(1.74)	(1.56)	(0.98)	(1.54)
Average number of distractors	2.12	8.45	3.01	2.06
(STDEV)	(0.86)	(0.88)	(0.72)	(1.58)
Total number of test items	1,071	2,604	28,457	29,206

Results of Stage 3: Coherent Item Model Item Generation Results

A total of 1,071 items were generated with three sets of distractors and more than one candidate. The item models with less than three distractor options were removed as they were not adequate for generating conventional MC test items. All 2,604 candidate texts were mapped with more than three sets of distractors and more than one option candidate using the second item model. Table 12 provides a list of example stems that were generated and introduced in the coherent item models. As specified in the item model, the stem or question statement was generated by identifying the topic keyword that best explains the given text based on their topic modelling results. Table 13 provides a list of example stems that were generated and introduced in the coherent item model 2. More specific examples of generated items from the coherent item model 1 and 2 are provided in Tables 14 and 15 and Appendix A1 and A2. In summary, a total of 3,675 items could be generated using the two coherent item models, which include more than one stem and more than three distractors as options.

Table 12

Example Stems Introduced from Coherent Item Model 1

Topic	Example Stems from Coherent Item Model 1
Topic 1	- The main character's feeling "hopeful" is most likely related to the statement: - The main character's feeling "excitement" is most likely related to the statement:
Topic 2	- The main character's feeling "cared" is most likely related to the statement: - The main character's feeling "cheerful" is most likely related to the statement:
Topic 3	- The main character's feeling "Terrible" is most likely related to the statement: - The main character's feeling "Miserable" is most likely related to the statement:
Topic 4	- The main character's feeling "trapped" is most likely related to the statement: - The main character's feeling "attacked" is most likely related to the statement:
Topic 5	- The main character's feeling "cared" is most likely related to the statement: - The main character's feeling "loved" is most likely related to the statement:
Topic 6	- The main character's feeling "furious" is most likely related to the statement: - The main character's feeling "irriatated" is most likely related to the statement:
Topic 7	- The main character's feeling "danger" is most likely related to the statement: - The main character's feeling "hurt" is most likely related to the statement:
Topic 8	- The main character's feeling "frightened" is most likely related to the statement: - The main character's feeling "terrified" is most likely related to the statement:
Topic 9	- The main character's feeling "welcomed" is most likely related to the statement: - The main character's feeling "amazed" is most likely related to the statement:
Topic 10	- The main character's feeling "cheering" is most likely related to the statement - The main character's feeling "lucky" is most likely related to the statement:

Table 13

Example Stems Introduced from Coherent Item Model 2

Topic	Example Stems from Coherent Item Model 1
Topic 1	What can be reasonably inferred from the passages that the main characters felt: - “Glad yeh found the place all righ'! We're doin' thestrals today — “
Topic 2	What can be reasonably inferred from the passages that the main characters felt: - “Mister Dursley, however, had a perfectly normal, owl-free morning.”
Topic 3	What can be reasonability inferred from the passage that the main characters felt: - "Marge's ill," he informed Aunt Petunia.', 'Oh my goodness — Vernon!'" - “Had Dumbledore actually cared about Harry at all? Or had Harry been nothing more than a tool to be polished and honed, but not trusted, never confided in?”
Topic 4	What can be reasonably inferred from the passages that the main characters felt: - “He heard Hermione's scream, Ron's yell, and a series of sickening metallic thuds, which told him that Xenophilius had been blasted off his feet and fallen backward down the spiral stairs.”
Topic 5	What can be reasonably inferred from the passages that the main characters felt: - "Oho! ‘One of my best friends is Muggle-born, and she's the best in our yeah' I'm assuming this is the very friend of whom you spoke, Harry?"
Topic 6	What can be reasonably inferred from the passages that the main characters felt: - “Nasty, common name, if you ask me. "Oh, yes," said Mister Dursley, his heart sinking horribly.” - “I know that," said Professor McGonagall irritably. "But that's no reason to lose our heads.”he main charater's feeling "Furious” is most likely related to the statement:
Topic 7	What can be reasonably inferred from the passages that the main characters felt: - “"Harry, could this be — ? Aargh!" Hermione screamed in pain, and Harry turned his wand on her in time to see a jeweled goblet tumbling from her grip. "It burned me!" moaned Hermione, sucking her blistered fingers.”
Topic 8	What can be reasonably inferred from the passages that the main characters felt: - “Aunt Petunia obviously scented danger, too, because she said quickly, "And we'll buy you another two presents while we're out today. Is that all right?” - “Hi, Harry! Wondered where you'd got to!" Hermione slid off the desk. "You shouldn't leave Lavender waiting outside," she said quietly.”

Topic 9	What can be reasonably inferred from the passages that the main characters felt: - “"Morning!" said Mister Weasley brightly. "Morning," said the Muggle.” - “Hello, Hagrid – Oh, it’s wonderful to see you two again – Are you coming into Gringotts, Harry?
Topic 10	What can be reasonably inferred from the passages that the main characters felt: - “I'll use the Invisibility Cloak," said Harry. "It's just lucky I got it back.”

Table 14

Coherent Topic Item Model 1 and the Example Generated Items

Coherent Topic Item Model 1

Input:
[Coherent toic text] [Topic A Sentences] [Topic A Keyword] [Topic A⁻ Sentences] [Characters]

Reading passage:
[Coherent topic text]

Stem:
The main character’s [Characters] feeling [Topic A Keyword] is most likely related to the statement:

Options:
Answer: [Topic A Sentences]
Distractors: [Topic A⁻ Sentences]

Example Generated Questions

The main characters’ (Harry, Hermione, Ron) feeling “scared” is most likely related to the statement:

A. **“No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper.”***
B. Harry and Ron went back to the common room. Harry had just said, “At least Hermione’s on Snapes’ tail.”
C. “Oh, come off it, you don’t think we’d let you go alone?” “Of course not” said Hermione briskly.
D. “I’m never going over to the Dark side! Voldemort killed my parents, remember?”

The main character's (Harry) feeling "welcomed" is most likely related to the statement:

A. **"Hello, Hagred – Oh, it's wonderful to see you two again – Are you coming into Gringotts, Harry?"***
B. "Sulkin' around Knockturn Alley, I dunni – dodgy place, Harry – don't want no one ter see yeh down there –"
C. "I'm staying with the Weasleys but we got separated," Harry explained.
D. "I should ruddy weel think not", growled Hagrid.

The main character's (Harry) feeling "fooled" is most likely related to the statement:

A. **"And why did he fake his death?" "Because he knew you were about to kill him like you killed my parents!"***
B. "He approached Lupin and the struggling rat, and his wet eyes suddenly seemed to be burning in his face."
C. "I persuaded Lily and James to change to Peter at the last moment, persuaded them to use him as Secret- Keeper instead of me"
D. "If he really is a rat, it won't hurt him."

The main character's (Harry) feeling "furious" is most likely related to the statement:

A. **"He had never before considered the possibility that there might be another teacher in the world but as he walked back toward Gryffindor Tower he had to admit he had found a contender."***
B. "She's evil, he thought, as he climbed a staircase to the seventh floor, she's an evil, twisted, mad, old"*
C. "Harry's third detention passed in the same way as the previous two, except that after two hours the words"
D. "I must not tell lies" did not fade from the back of Harry's hand, but remained scratched there, oozing droplets of blood

The main character's (Ron) feeling "terrified" is most likely related to the statement:

A. **"Harry! Ron was standing over him looking extremely frightened. "He's really ill," said a scared voice."***
B. "It wasn't a dream. Not an ordinary dream. I did it. He could hear Seamus and Dean muttering but did not care."
C. "You did promise her, you know, Harry. I think you'd better give her something else instead. How about your Firebolt? "
D. "Your dad! He's been bitten, it's serious, there was blood everywhere."

The main character's (Harry's family) feeling "painful" is most likely related to the statement:

A. **"Not Harry, please no, take me, kill me instead — "***
B. “And his scream was Harry's scream, his pain was Harry's pain .",
C. “And now he stood at the broken window of Bathilda's house, immersed in memories of his greatest loss, and at his feet the great snake slithered over broken china and glass.”
D. “He did not like it crying, he had never been able to stomach the small ones whining in the orphanage — "

The main charater's (Harry) feeling "guilty" is most likely related to the statement:

A. “**Nobody’s ever asked me to a party before, as a friend!”***
B. “that was a mistake. I'll get Hermione to put it right for me”
C. “And sure enough, in no time at all the whole school seemed to know that Harry Potter was taking Luna Lovegood to Slughorn’s party.”
D. “Harry tried to feel pleased that Ginny was glad he was taking Luna to the party, but could not quite manage it.”

Note. * answer keys

Table 15

Coherent Topic Item Model 2 and the Example Generated Items

Coherent Topic Item Model 2

Input: [Coherent toic text] [Topic A Sentences] [Topic A Keyword] [Topic A⁻ Sentences] [Characters]

Reading passage:
[Coherent topic text]

Stem:
What can be reasonably inferred from [Topic A Sentences] that the main character [Characters] felt:

Options:
Answer: [Topic A Keyword]
Distractors: [Topic A⁻ Keywords]

Example Generated Questions

What can be reasonably inferred from line 19 of the passage " "I know that," said Professor McGonagall irritably.'"But that's no reason to lose our heads." that the character (Professor McGonagall) felt:

A. Annoyed*	D. Frightened	G. Guilty
B. Worried *	E. Terrified	H. Hatred
C. Unpleasant*	F. Struggled	

What can be reasonably inferred from line 2-3 of the passage " "In the car crash when your parents died," she had said. " And don't ask questions." " that the main character (Harry) felt:

A. Curious*	D. Terrified	G. Danger
B. Ignored*	E. Rage	H. Scared
C. Discouraged*	F. Temper	

What can be reasonably inferred from line 16-17 of the passage "Aunt Petunia obviously scented danger, too, because she said quickly, "And we'll buy you another two presents while we're out today. Is that all right?" " that that the character (Aunt Petunia) felt:

A. Alarmed*	D. Scared	G. Generous
B. Worried*	E. Danger	H. Hopeful
C. Terrified*	F. Relieved	

What can be reasonably inferred from line 14-15 of the passage " "Had Dumbledore actually cared about Harry at all? Or had Harry been nothing more than a tool to be polished and honed, but not trusted, never confided in?" " that that the main character (Harry) felt:

A. Furious*	D. Frightened	G. Excited
B. Rage*	E. Loved	H. Attacked
C. **Hatred***	F. Destoryed	

What can be reasonably inferred from line 3-4 of the passage ""Hufflepuff'll have to lose by at least two hundred points," said George. " " that the character (Geroge) felt:

A. Hopless*	D. Lucky	G. Delighted
B. Doubtfu*l	E. Excited	H. Scared
C. **Losing***	F. Clever	

What can be reasonably inferred from line 13-16 of the passage "Tell us about being attacked by the giants and Harry can tell you about being attacked by the dementors — " Hagrid choked in his mug and dropped his steak at the same time; a large quantity of spit, tea, and dragon blood was sprayed over the table as Hagrid coughed and spluttered and the steak slid, with a soft splat, onto the floor." That the character (Hagrid) felt:

A. Horrified*	D. Disappointed	G. Awful
B. Impatient*	E. Hopeless	H. Excited
C. Surprised*	F. Mad	

What can be reasonably inferred from line 24-15 of the passage "There was a pause during which Harry stared fixedly at a large dead frog suspended in a purple liquid in its jar." That the character (Harry) felt:

A. Forbidden*	D. Impatient	G. Lost
B. Frightened*	E. Feared	H. Drowned
C. Trapped*	F. Miserable	

What can be reasonably inferred from line 6-8 of the passage "It's the most powerful love potion in the world!" said Hermione. "Quite right! You recognized it, I suppose, by its distinctive mother-of-pearl sheen?" That the character (Hermione) felt:

A. Excited*	D. Impatient	G. Punished
B. Enjoying*	E. Comfortable	H. Rushed
C. **Interested***	F. Powerful	

Note. * answer keys

Results of Stage 3: Divergent Item Model Item Generation Results

A total of 28,457 items were generated with three sets of distractors and more than one candidate stem using the divergent item model 1. Similarly, a total of 29,206 test items were generated with the acceptable number of distractors (≥3) and the stem (≥1) using the second divergent item model. The test items generated from these item models directly highlighted parts of the text to assess the underlying sentiment of the given text. Tables 16 and 17 provide a list of example items generated from the coherent item models. More specific examples of generated

items are provided in Appendix A3 and A4. In summary, a total of 57,663 items could be generated using the two coherent item models using the source texts of 28,457. All of the generated test items included more than one stem and more than three distractors as options to satisfied the conventional MC test item writing guidelines.

Table 16

Divergent Topic Item Model 1 and the Example Generated Items

Divergent Topic Item Model 1

Input: [Divergent toic text] [Topic A Sentences] [Topic B Sentences] [Topic A Keyword] [Topic B Keyword] [Topic AB⁻ Keywords]=

Reading passage:
[Divergent topic text]

Stem:
How is the main character's sentiment described in [Topic A sentence] different from [Topic B sentences]?

Options:
Answer: [Topic A keywords] and [Topic B keywords]
Distractor: [Topic A keywords] and [Topic AB⁻ keywords]
Distractor: [Topic A keywords] and [Topic AB⁻ keywords]
Distractor: [Topic AB⁻ keywords] and [Topic B keywords]
Distractor: [Topic AB⁻ keywords] and [Topic B keywords]

Example Generated Questions

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "delighted" in (A) while the character felt "confused" in (B).*
B. The character felt "amused" in (A) while the character felt "cheerful" in (B).
C. The character felt "lucky" in (A) while the character felt "hopeful" in (B).
D. The character felt "treated" in (A) while the character felt "suspicious" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "trusting" in (A) while the character felt "suspicious" in (B).*
B. The character felt "panicked" in (A) while the character felt "hopeful" in (B).
C. The character felt "impatient" in (A) while the character felt "excited" in (B).
D. The character felt "satisfied" in (A) while the character felt "interested" in (B)

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "delighted" in (A) while the character felt "suspicious" in (B).*
B. The character felt "uncomfortable" in (A) while the character felt "excited" in (B).
C. The character felt "frightened " in (A) while the character felt "hopeful" in (B).
D. The character felt "enthusiastic" in (A) while the character felt "terrified" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "confused" in (A) while the character felt "doubtful" in (B).*
B. The character felt "amazed" in (A) while the character felt "confused" in (B).
C. The character felt "relieved" in (A) while the character felt "lucky" in (B).
D. The character felt "bored" in (A) while the character felt "powerless" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "amazed" in (A) while the character felt "determined" in (B).*
B. The character felt "excited" in (A) while the character felt "scared" in (B).
C. The character felt "feared" in (A) while the character felt "angry" in (B).
D. The character felt "relieved" in (A) while the character felt "disappointed" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "rage" in (A) while the character felt "cheerful" in (B).*
B. The character felt "destroyed" in (A) while the character felt "amused" in (B).
C. The character felt "disappointed" in (A) while the character felt "scared" in (B).
D. The character felt "painful" in (A) while the character felt "relieved" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "convinced" in (A) while the character felt "regret" in (B).*
B. The character felt "mad" in (A) while the character felt "disappointed" in (B).
C. The character felt "satisfied" in (A) while the character felt "forbidden" in (B).
D. The character felt "rage" in (A) while the character felt "painful" in (B).

How is the main characters sentiment described in (A) and (B) different?

A. The character felt "enjoying" in (A) while the character felt "furious" in (B).*
B. The character felt "suspicious" in (A) while the character felt "rage" in (B).
C. The character felt "fun" in (A) while the character felt "weak" in (B).
D. The character felt "enthusiastic" in (A) while the character felt "feared" in (B).

Note. * answer keys

Table 17

Divergent Topic Item Model 1 and the Example Generated Items

Divergent Topic Item Model 2

Input:
[Divergent toic text] [Topic A Sentences] [Topic B Sentences] [Topic C Sentences] …

Reading passage:
[Divergent topic text]

Stem:
Which of the following statements indicates different sentiment from the others?

Options:
Answer: [Topic A Sentences] & Distractor: [Topic A⁻ Sentences]
Answer: [Topic B Sentences] & Distractor: [Topic B⁻ Sentences]
Answer: [Topic C Sentences] & Distractor: [Topic C⁻ Sentences]

Example Generated Questions

Which of the following indicates different sentiment of the characters from the others?

A. "The Potters, that's right, that's what I heard — yes, their son, Harry — "*
B. "Mister Dursley, however, had a perfectly normal, owl-free morning."
C. "Little tyke," chortled Mister Dursley as he left the house."
D. "Mister Dursley hummed as he picked out his most boring tie for work, and Miss Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair."

Which of the following indicates different sentiment of the characters from the others?

A. "Harry didn't much like Peeves, but couldn't help feeling grateful for his timing." *
B. “Filch roared, flinging down his quill in a transport of rage.”
C. "It's only a bit of mud to you, boy, but to me it's an extra hour scrubbing!" shouted Filch, a drip shivering unpleasantly at the end of his bulbous nose.
D. Dabbing at his streaming nose, Filch squinted unpleasantly at Harry, who waited with bated breath for his sentence to fall.

Which of the following indicates different sentiment from the others?

A. “So clever, the way you trapped that last one with the tea-strainer —”
B. “Well, I'm sure no one will mind me giving the best student of the year a little extra help,”
C. “Yes, nice, isn't it?" he said, misreading the revolted look on Ron's face. “I usually save it for book signings.”
D. “Hermione put it carefully into her bag and they left, trying not to walk too quickly or look too guilty.” *

Which of the following indicates different sentiment from the others?

A. “Dursleys had always forbidden questions about his wizarding relatives.”
B. “Ah, said a nasty little voice in his brain, but the Sorting Hat wanted to put you in Slytherin, don't you remember?”
C. “Harry got up and left through the portrait hole, wondering where Justin might be. Shivering,”
D. “Hannah, said the stout boy solemnly, he's a Parselmouth. They called Slytherin himself Serpent-tongue.” *

Which of the following indicates different sentiment from the others?

A. “But — only a Gryffindor could have stolen — nobody else knows our password —”
B. “Harry had been staring down the packed Gryffindor table, wondering if the new owner of Riddle's diary was right in front of his eyes.”
C. “He'd have to tell a teacher all about the diary, and how many people knew why Hagrid had been expelled fifty years ago?”
D. "Perfect Quidditch conditions!, said Wood enthusiastically” *

Which of the following indicates different sentiment from the others?

A. “Not a punishment, Hagrid, more a precaution. If someone else is caught, you'll be let out with a full apology —”
B. “Mister Lucius Malfoy strode into Hagrid's hut, swathed in a long black traveling cloak, smiling a cold and satisfied smile.”
C. “My dear man, please believe me, I have no pleasure at all in being inside your — er — d'you call this a house?”
D. “Already here, Fudge," he said approvingly. "Good, good. ”*

Which of the following indicates different sentiment from the others?

A. “When Harry got outside again, he found Ron being violently sick in the pumpkin patch.”*
B. “I'll never forgive Hagrid. We're lucky to be alive.”
C. “That's exactly Hagrid's problem!" said Ron, thumping the wall of the cabin.”
D. "He always thinks monsters aren't as bad as they're made out, and look where it's got him! A cell in Azkaban!"

Which of the following indicates different sentiment from the others?

A. “I wouldn't mind knowing how Riddle got an award for special services to Hogwarts either.”
B. “Maybe he murdered Myrtle; that would've done everyone a favor.”
C. “Oooh, it might have hidden powers,” said Hermione”
D. “He never wrote in it,” said Harry” *

Note. * answer keys

Item Generation Process Validation

Validation of the item generation process enables an evidence-based approach to control for test item quality. In my AIG framework, test items are constructed as the final product of carefully integrated three-stage item development process. Hence, I evaluated the outcomes of each item generation stage in terms of their consistency. Using this approach, I could ensure that the final product (i.e., test items) is of the expected format, content, and quality. More specifically, the evaluation of the first stage focuses on subtopic keywords and their semantic categories. The second stage validation focuses on the sentimental alignment between topic

keywords and topic sentences. The third state focues on the item generation procedure by visualizing and applying the inference test question generation guidelines proposed by Ennis (1969, 1973, 1981), Collins, Brown, and Larkin (1980), and Phillips (1989).

Validation of Stage 1: Text to Topic Results Validation. The first stage of the item generation process focused on extracting sentiment subtopics from text. The quality of the sentiment-weighted topic modelling results required evaluation for its appropriateness of the topic keywords. This evaluation was to ensure the topic keywords belonged to a suitable semantic category for item generation introduced in this dissertation. In order for the topic keywords to serve as part of the item generation process, the keywords should represent sentimental aspects stemming from the interaction between the characters in the given text. Hence, I adopted the approach of assessing topic hierarchy using WordNet to evaluate the lexical taxonomy of topic keywords (e.g., Monteiro Vieira & Brey, 2012). WordNet is a large lexical resource in English, which provides information about words and their semantic relationships (Miller, 1995). WordNet provides detailed information about lexical taxonomy by disambiguating the semantic associations and relationships between words. The relationship between words is presented using hypernymy or general and overarching terms and hyponymy or specific instance. For example, the word “colour” would be a hypernym, or general term of the word “red”, “blue”, and “yellow”, representing specific instances of the general term.

In terms of the general “hypernym” associated with the interactions and the sentimental value, WordNet introduces the word “abstraction” as “a general concept formed by extracting common features from specific examples” (Miller, 1995). This encompasses five specific concepts or hyponyms, which are of the current analysis interest, such as “attributes”, “relationships”, “psychological features”, “communication”, and “group” as its lower-taxonomy

words (Table 18). Hence, the validation of the semantic appropriate of the topic keywords was conducted by investigating whether the general semantic categories of the topic keywords belong to any of the five specific concepts in Table 18.

Table 18

Synsets Related to the Conception of "Abstraction" according to WordNet

Taxonomy	WordNet Definitions	Hyponyms or Lower rank words:
group.n.01	Any number of entities (members) considered as a unit.	"association.n.02", "community.n.06", "people.n.01", "kingdom.n.06", "social_group.n.01", "biological_group.n.01", "ethnic_group.n.01"
communication. n.02	Something that is communicated by or to or between people or groups.	"auditory_communication.n.01", " expressive_style.n.01", "message.n.01", "psychic_communication.n.01", "signal.n.01"
psychological feature.n.01	A feature of the mental life of a living organism.	"cognition.n.01", "event.n.01", "motiavation.n.01"
attribute.n.02	An abstraction belonging to or characteristic of an entity.	"human_nature.n.01", "cheerfulness.n.01", "character.n.09", "personality.n.01"
relation.n.01	An abstraction belonging to or characteristic of two entities or parts together.	"social_relation.n.01", "opposition.n.02", "unconnectedness.n.01", "kinship.n.02", "relationship.n.01", "connection.n.01"

The evaluation of the taxonomical appropriateness of the topic keywords was conducted on the top 20 vocabularies representing each topic with the highest contributing weights. The taxonomy disambiguation of the words was only limited to nouns, hence, the words with other parts-of-speech, such as adverbs and adjectives, could not be further investigated. The results indicated that all of the nouns that could be disambiguated using the WordNet taxonomy were classified as one of the five target taxonomy related to sentiment and interaction (Table 19). For instance, all keywords in Topic 10 were classified as one of the taxonomies related to abstraction (e.g., group, communication, psychological feature, attribute, relation). Eleven out of twenty disambiguated words belonged to the lexical taxonomy related to "attributes", which is defined

as “an abstraction belonging to or characteristic of an entity”, such as “human nature”, “personality”, and “characteristics”. This taxonomy, in fact, included the majority of the keywords across the ten topics. Topic 1 and 8 included a relatively smaller number of keywords that could be evaluated. Yet, all nouns in these topic keywords were classified as one of the target lexical taxonomy categories (Table 19).

To summarize, the taxonomy disambiguation results indicated that the topic keywords were generated from the target domain (e.g., character’s sentiment and interaction). This, in turn, identifies that the sentiment-weighted topic keywords were generated from the appropriate lexical taxonomy to represent the sentimental and interactional aspects of the stories. The validation evidence provides important theoretical and empirical support for one of the key item construction components in this dissertation, the topic keywords.

Table 19

Topic Keyword Lexical Taxonomy Disambiguation Results

	Topic									
Hypernym Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
group.n.01	0	2	1	0	1	1	0	0	0	2
communication.n.02	1	2	0	0	1	2	1	1	1	3
psychological feature.n.01	1	2	3	5	3	2	3	1	3	3
attribute.n.02	4	4	5	9	6	6	5	3	6	11
relation.n.01	0	0	0	0	0	0	1	0	0	1
Total Vocabulary	6	10	9	14	11	11	10	5	10	20
Total N Nouns	6	10	9	14	11	11	10	5	10	20
(%) / Total N Nouns	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
(%) / Top 20 Vocabs	30%	50%	45%	70%	55%	55%	50%	25%	50%	100%

Validation of Stage 2: Subtext Topic Sentence and Keyword Representation

Validation. The second stage of the item generation process focused on filtering and categorizing text based on their subtopic structure that are identified in Stage 1. One important outcome of this process was the identification of key topic sentences. The validity of the key sentences representing the topic structures in the subtexts requires evaluation. This evaluation is necessary to ensure the alignment between the final input (i.e., topic keywords and sentences) to the item models in Stage 3. The item models require topic key sentences and keywords to have a strong association to present equivalent or similar sentimental topics. In order to evaluate their alignment, the sentimental values of the key sentences from each candidate subtext and the topic keywords were computed using VADER (see page 35 in Chapter 3; Hutto & Gilbert, 2014). The correlation between the sentimental representation between the two components was, then, computed and evaluated in terms of their topic representation.

Table 20 provides a correlation between the key sentences and keywords extracted from the coherent topic subtexts. A total of 2,604 coherent topic subtexts were categorized based on their dominant topics (i.e., Topic 1 – Topic 10). Next, the positive and negative sentiment scores of the topic keywords and key sentences were compared. The results indicate that the topic keywords and the sentences showed moderate to high correlation coefficients in both positive and negative sentiment scores. This result suggests that the two elements were aligned consistently to communicate similar or equivalent sentimental values in the coherent topic subtexts. The sentimental alignment between the two elements ensures the item models used to generate test items will produce consistent context.

Table 20

A Correlation between the Key Sentences and Keywords of Coherent Topic Subtexts

Pearson's Correlation		**Sentiment Score**	
		Positive	**Negative**
Topic 1 Keywords	Topic 1 Key sentences	0.36	0.83
Topic 2 Keywords	Topic 2 Key sentences	0.22	0.32
Topic 3 Keywords	Topic 3 Key sentences	0.13	0.51
Topic 4 Keywords	Topic 4 Key sentences	0.47	0.36
Topic 5 Keywords	Topic 5 Key sentences	0.35	0.52
Topic 6 Keywords	Topic 6 Key sentences	0.48	0.31
Topic 7 Keywords	Topic 7 Key sentences	0.45	0.29
Topic 8 Keywords	Topic 8 Key sentences	0.68	0.52
Topic 9 Keywords	Topic 9 Key sentences	0.51	0.18
Topic 10 Keywords	Topic 10 Key sentences	0.31	0.42
Total Keyworrds	Total Topic Key Sentences	0.48	0.46

Tables 21 and 22 provide correlation coefficients between the key sentences and keywords extracted from the divergent topic subtexts. A total of 26,939 divergent topic subtexts were categorized based on their dominant topics and the second dominant topics. The results indicated that the topic keywords and the sentences showed a relatively high correlation in both positive and negative sentiment scores. The high correlation between the keywords and sentences suggest that the two elements were aligned in a consistent manner to communicate similar or equivalent sentimental values in the divergent topic subtexts. Hence, the two elements could be interchangeably used to generate test items, which require examinees to evaluate their alignment .

Table 21

A Correlation between the Key Sentences and Keywords of Divergent Topic Subtexts 1

Pearson's Correlation		**Sentiment Score**	
		Positive	**Negative**
Topic 1 Keywords	Topic 1 Key sentences	0.31	0.41
Topic 2 Keywords	Topic 2 Key sentences	0.30	0.35
Topic 3 Keywords	Topic 3 Key sentences	0.35	0.37
Topic 4 Keywords	Topic 4 Key sentences	0.41	0.34
Topic 5 Keywords	Topic 5 Key sentences	0.40	0.37
Topic 6 Keywords	Topic 6 Key sentences	0.46	0.36
Topic 7 Keywords	Topic 7 Key sentences	0.37	0.30
Topic 8 Keywords	Topic 8 Key sentences	0.36	0.35
Topic 9 Keywords	Topic 9 Key sentences	0.35	0.40
Topic 10 Keywords	Topic 10 Key sentences	0.32	0.37
Total Keyworrds	Total Topic Key Sentences	0.39	0.41

Table 22

A Correlation between the Key Sentences and Keywords of Divergent Topic Subtexts 2

Pearson's Correlation		**Sentiment Score**	
		Positive	**Negative**
Topic 1 Keywords	Topic 1 Key sentences	0.35	0.40
Topic 2 Keywords	Topic 2 Key sentences	0.33	0.37
Topic 3 Keywords	Topic 3 Key sentences	0.38	0.40
Topic 4 Keywords	Topic 4 Key sentences	0.37	0.38
Topic 5 Keywords	Topic 5 Key sentences	0.36	0.37
Topic 6 Keywords	Topic 6 Key sentences	0.43	0.37
Topic 7 Keywords	Topic 7 Key sentences	0.42	0.41
Topic 8 Keywords	Topic 8 Key sentences	0.42	0.44
Topic 9 Keywords	Topic 9 Key sentences	0.46	0.43
Topic 10 Keywords	Topic 10 Key sentences	0.37	0.40
Total Keyworrds	Total Topic Key Sentences	0.40	0.41

Validation of Stage 3: Item Generation Results. The previous validation results demonstrated the quality and alignment between the integral components and the input of item modelling – topic keywords and key sentences. The second source of validation showed that the sentiment-weighted topic modelling results could produce the topic keywords of appropriate categories closely associated with sentiment and interaction. The categories indicated that all topic keywords that are nouns could be explained using the abstract entity. The categories included “group”, “communication” “psychological feature”, “attribute”, and “relation”. Likewise, validation of Stage 3 demonstrated that the topic keywords and sentences produced from the topic modelling results have moderate to high correlations in terms of their sentimental representation. The sentimental alignment between the topic keywords and the key sentences is important to ensure the quality of the generated items. This is because the item models are designed to use the association between the topic keywords and sentences to assess the examinee’s knowledge to make correct inferences about the character’s interactions and sentiment from the given text.

The comparison to the inference item review guideline emphasizes the importance of the alignment between the two elements to ensure item quality. Extensive reviews of the item quality guidelines by Ennis (1969, 1973, 1981), Collins, Brown, and Larkin (1980) and Phillips (1989) introduce comprehensive evaluative guidelines of inference questions in reading comprehension tests. The guidelines by Ernnis (1969, 1973, 1981), Collins et al. (1980) and Phillips (1989) suggest that well-constructed inference items should include the reading passage and the item element that are presenting adequate amount and quality of information required for the inferencing situation, accurate or complete with adequate evidence for asserted information, relevant to the ongoing situation, and unambiguous and clear.

Under the item modelling conditions presented in this dissertation, the criteria are satisfied based on the following points. First, the item models ensure that the examinees are not provided with too little or too excessive amount of information by systematically providing only integral parts of the information. This information could be a sentence, a keyword, or a sentence and a keyword depending on the specific item model and the text type.

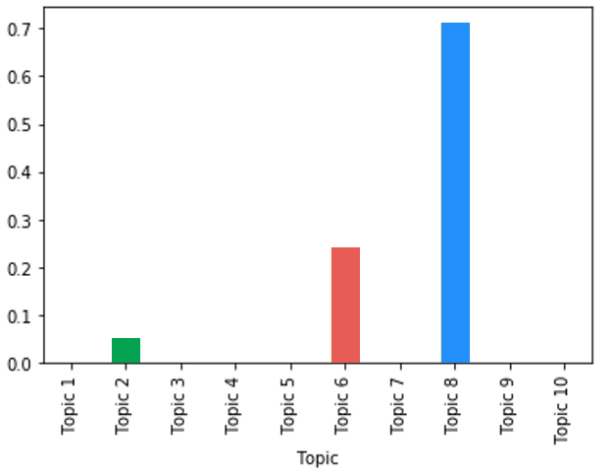
Second, item models produce test items using the input of topic keywords and topic key sentences. These elements are directly located and extracted from the given text to provide accurate and complete evidence about the overall sentimental aspects of the given text. Moreover, the strong correlation between the two elements in terms of their sentiment representation ensures that only adequate types of information are presented and manipulated to generate variations of test items.

Third, the item models identify plausible-but-incorrect options, or distractors, by locating the topic keywords and key sentences of the topic structure that are relatively local or less dominant in the given text. This ensures that the distractor elements in item generation contain consistent and relevant types of information, in our case, textual evidence indicating certain sentimental values.

To ensure that the item generation does not involve any ambiguous information in their item model component, topic keywords and sentences with ambiguous underlying sentimental structure is detected and removed. In the next section, I demonstrate how the item quality can be evaluated by elaborating on the example in Figures 21 and 22.

"Oh, shut up," said Hermione, but she agreed to go and watch out for Snape. "And we'd better stay outside the third-floor corridor," Harry told Ron. "Come on. " No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper. "I suppose you think you're harder to get past than a pack of enchantments!" "Enough of this nonsense! If I hear you've come anywhere near here again, I'll take another fifty points from Gryffindor! Yes, Weasley, from my own House!" Harry and Ron went back to the common room. Harry had just said, "At least Hermione's on Snape's tail," when the portrait of the Fat Lady swung open and Hermione came in. "I'm sorry, Harry!" "Snape came out and asked me what I was doing, so I said I was waiting for Flitwick, and Snape went to get him, and I've only just got away, I don't know where Snape went. " "Well, that's it then, isn't it?" Harry said. "I'm going out of here tonight and I'm going to try and get to the Stone first. " "You're mad!" said Ron. "You can't!" said Hermione. "After what McGonagall and Snape have said? You'll be expelled!" "SO WHAT?" Harry shouted. "Don't you understand? If Snape gets hold of the Stone, Voldemort's coming back! There won't be any Hogwarts to get expelled from! D'you think he'll leave you and your families alone if Gryffindor wins the House Cup? If I get caught before I can get to the Stone, well, I'll have to go back to the Dursleys and wait for Voldemort to find me there, it's only dying a bit later than I would have, because I'm never going over to the Dark Side! Voldemort killed my parents, remember?" "You're right, Harry," said Hermione in a small voice. "I'll use the Invisibility Cloak," said Harry. "It's just lucky I got it back. " "But will it cover all three of us?" said Ron. "All — all three of us?" "Oh, come off it, you don't think we'd let you go alone?" "Of course not," said Hermione briskly.

Figure 21. Example subtext candidate with topic sentences highlighted (blue=Topic 8, red=Topic 6, purple=Topic 6 & Topic 8, green=Topic 2).



Topic 2	Topic 6	Topic 8
Party	Fury	Terrified
Saved	Anger	Worried
Love	Rage	Scared
Lucky	Destroy	Frightened
Hopeful	Fake	Nasty
Cheerful	Forbidden	Dangerous
Enjoying	Stolen	Guilty
Hopefully	Irritated	Trapped
Interested	Terrified	Danger
Delighted	Scream	Fooled
Perfectly	Furious	Weak

Figure 22. Example subtext candidate with topic keywords.

Figure 21 provides an example subtext candidate from the coherent topic category. The subtext provides Topic 8 as a dominant topic structure (0.72 or 72%) with additional evidence from Topic 6 (0.26 or 26%) and Topic 2 (0.02 or 2%). Figure 22 provides the topic keywords of the three topics (Topic 8, 6, and 2) shown in the example text. On average, the topic keywords and topic sentences in Topic 8 produced a moderate sentimental correlation in a coherent subtext (positive sentiment correlation coefficient= 0.68, negative= 0.52). This demonstrates the close

associations between the highlighted Topic 8 key sentences (“No sooner had they read the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper”; “You’re mad!” said Ron. “You can’t!” said Hermione. “After what McGonagall and Snape had said? You’ll be expelled”; Figure 21) and the topic keywords (“Terrified”, “Worried”, “Scared”, and “Frightened”; Figure 22).

Using the topic sentence and topic keyword components, Table 23 presents the list of the item model elements – stems, keyed-option, and distractors – which are eligible to generate test items from the given coherent subtext. Four stems could be generated with the relevant keywords from Topic 8, “scared”, “terrified”, “worried”, and “frightened”. Similarly, three possible keyed options are presented from the Topic 8 sentences, while the distractors are extracted from the key sentences from different topics (Topic 6 and Topic 2). In selecting appropriate distractors, the topic sentences representing overlapping topics, such as Topic 6 and Topic 8, were removed to ensure the quality of distractors (Invalid options in Table 23). While these options provide some mixing sentiment of Topic 6 and Topic 8 according to the results, they cannot be presented together as valid options as they might introduce some unintentional ambiguity in the test items. Hence, item modeling addressed those cases where the topic sentences present some mixing sentiment by removing them from item generation. Random combinations of the three components could result in a total of 4 (stem) * 3 (keyed-option) * 4 (3 distractors) = 48 variations of a 4-option multiple-choice question. Table 24 provides a list of example combinations of item elements generated as final 4-option multiple-choice questions.

Table 23

Item Components Generated from the Example Subtext in Figure 21

	Item Component	Topic
Stem	The main character's feeling "scared" is most likely related to the statement:	8
	The main character's feeling "terrified" is most likely related to the statement:	8
	The main character's feeling "worried" is most likely related to the statement:	8
	The main character's feeling "frightened" is most likely related to the statement:	8
Keyed Option	"No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper."	8
	"You're mad!" said Ron. "You can't!" said Hermione.	8
	"After what McGonagall and Snape have said? You'll be expelled"	8
Distractors	"I suppose you think you're harder to get past than a pack of enchantments!"	6
	"Enough of this nonsense!"	6
	"If I get caught before I can get to the Stone, wellm I'll have to go back to the Dursleys and wait for Voldemort to find methere, its' only dying a bit later than I would have, because I'm never going over to the Dark Side"	6
	"It's just lucky I got it back."	2
Invalid Option	"If Snape gets hold of the Stone, Voldemort's coming back!"	6+8
	"There won't be any Hogwarts to get expelled from!"	6+8
	"D'you think he'll leave you and your families alone if Griffindor wins the House Cup?"	6+8

Table 24

Example Combinations of 4-option MC questions from Table 23

The main characters' feeling "scared" is most likely related to the statement:

A. "No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper." *

B. "I suppose you think you're harder to get past than a pack of enchantments!"

C. "If I get caught before I can get to the Stone, wellm I'll have to go back to the Dursleys and wait for Voldemort to find methere, its' only dying a bit later than I would have, because I'm never going over to the Dark Side"

D. "It's just lucky I got it back."

The main characters' feeling "scared" is most likely related to the statement:

A. "You're mad!" said Ron. "You can't!" said Hermione. *

B. "I suppose you think you're harder to get past than a pack of enchantments!"

C. “If I get caught before I can get to the Stone, wellm I’ll have to go back to the Dursleys and wait for Voldemort to find methere, its’ only dying a bit later than I would have, because I’m never going over to the Dark Side”
D. “It’s just lucky I got it back.”

The main characters’ feeling “terrified” is most likely related to the statement:
A. “**No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper.” ***
B. “I suppose you think you’re harder to get past than a pack of enchantments!”
C. “If I get caught before I can get to the Stone, wellm I’ll have to go back to the Dursleys and wait for Voldemort to find methere, its’ only dying a bit later than I would have, because I’m never going over to the Dark Side”
D. “It’s just lucky I got it back.”

The main characters’ feeling “terrified” is most likely related to the statement:
A. “After what McGonagall and Snape have said? You’ll be expelled” *
B. “I suppose you think you’re harder to get past than a pack of enchantments!”
C. “If I get caught before I can get to the Stone, wellm I’ll have to go back to the Dursleys and wait for Voldemort to find methere, its’ only dying a bit later than I would have, because I’m never going over to the Dark Side”
D. “It’s just lucky I got it back.”

The main characters’ feeling “worried” is most likely related to the statement:
A. “No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper.” *
B. “I suppose you think you’re harder to get past than a pack of enchantments!”
C. “If I get caught before I can get to the Stone, wellm I’ll have to go back to the Dursleys and wait for Voldemort to find methere, its’ only dying a bit later than I would have, because I’m never going over to the Dark Side”
D. “It’s just lucky I got it back.”

Chapter Summary

Chapter 4 provided the results and the findings from the three primary stages of the AIG framework. The findings from the first stage indicated that the sentiment-weighted topic modelling approach could be used to generate distinct types of topic keywords to represent the content from the given texts. The second stage described how the rule-based parsing using topic

key sentences and topic weights resulted in a total of 2,604 and 28,475 coherent and divergent topic subtexts. The third stage of item generation with four-item models resulted in a total of 1,071, 2,604, 28,457, and 29,206 test items. The validation results of the sentiment-weighted topic keywords demonstrated that the keywords were generated from sentiment- and interaction-relevant lexical categories based on their semantic associations. Then, I evaluated whether the topic keywords and sentences could provide similar sentimental representation to assess their alignment for item generation. The qualitative evaluation using a comprehensive theoretical framework of inference item review revealed that the item development condition controlled and introduced in this AIG framework could ensure inference test items with desirable characteristics and, hence, adequate quality.

Chapter 5: Discussion

Chapter 5 focuses on restating the purpose of the study and discussing the implications of the findings presented in Chapter 4. The three stages of sentiment-topic modelling, rule-based parising, and item modelling are further discussed based on their outcomes and the methodological and empirical implications for item developers and educators. The chapter concludes by identifying and addressing the limitation of this disseratation followed by the directions of future research to overcome the limitation.

Purpose of the Study

Generating test items automatically has been a methodologically and practically important line of research to maximize the potential of technological innovation in educational assessment to resolve the previous challenges in item devleopment. The increased need and capacity of educational assessment using computer-based assessment had a dramatic impact on item development practices. The traditional approach of writing test items could not sufficiently support the exponentially growing demand for items using computerized assessments. Subjectivity and scalability were the two fundamental problems limiting the item creation capacity in the traditional processes. For instance, the traditional method of item writing relied on subject matter experts (SMEs) to write individual test items. This indicates that the SMEs were responsible for the full cycle of item generation, which includes writing, editing, reviewing, and revising each test item. Furthermore, the process was often conducted solitarily by individual SMEs. The practice of peer-review and support was ideal but relatively uncommon. Moreover, because of such practices, item writing heavily depended on the SME's understanding of the knowledge and skills within a specific content area.

The issues in scalability and subjectivity made the item writing practices less cost- and time-efficient and more subjective and error-prone (Rudner, 2010; Rush, Rankin, & White, 2016; Masters et al., 2001). To overcome these limitations, researchers from different disciplines have approached the problem using their unique methods. For example, researchers in educational assessment, measurement, and psychology have strived to create a viable and effective framework using the template-based approach. This approach was used to structure and convey the content knowledge of SMEs effectively and efficiently in order to scale the item generation process and lower item creation costs. This led to the state-of-the-art foundation of template-based approach for generating test items (Gierl & Lai, 2013; 2016; Gierl & Haladyna, 2012). Computer scientists, by way of comparison, have addressed the item generation problem with the focus on directly modelling and extracting test item content. For instance, important contextual information was extracted and modelled from the source text and documents (e.g., textbook, news article, Wikipedia) to restructure and map the content knowledge in the form of test items (e.g., Brown et al., 2005; Chen et al., 2006; Du et al., 2017; Gao et al., 2018; Mazidi, 2017; Narayan et al., 2020). These methods aimed to replace the manual knowledge structuring and item generation, which largely depended on pre-defined knowledge.

The development and introduction of the two frameworks demonstrate clear and promising paths for future item development paradigms in educational assessment. This reflects the trend for decreasing the amount of time required for manual content modelling by SMEs while at the same time increasing the number of the generate test items. Despite the promises, both the template-based and the non-template-based approachs in item generation have inherent problems. Template-based item generation could readily generate test items adhering to the operational administration guidelines and standards. However, the generation process heavily

depended on the ability to accurately model the knowledge structure by SMEs. This could be a highly complex process for certain domains, where extracting and communicating the decision-making processes is largely inexplicit. For example, reading comprehension item generation, which focuses on examinees' inferential and evaluation knowledge of the content, was often perceived as a daunting and challenging domain for automated item generation.

Non-template-based AIG item generation overcomes some of the problems by focused on explicitly modelling the knowledge contents from the source text. This was conducted by using the semantic or semantic features or by directly extracting and learning the integral parts of the text using sequential modelling approaches. Without any specific item generation guideline and theory, however, the generated items often suffered from a lack of quality for operational uses. To overcome the limitations, I proposed a hybrid approach for my dissertation research that integrated non-template and template-based item generation methods in the challenging domain of reading inference-type questions used in reading comprehension exams.

My proposed AIG method focused on extending the capacity of the previous template-based approach to generate quality items satisfying the operational administration standards. My framework uses advanced natural language processing techniques adopted and improved from the previous non-template-based approaches. The outcomes from my research demonstrate the capacity and the useability of the extended AIG framework to create inference-type questions from the Harry Potter series. The generated test items focused on assessing the examinee's inferential knowledge. My proposed AIG method focused on generating test items that target students' understanding in making correct inferences regarding the character's interaction and sentiment. The generated test items, thereby, represented a large pool of test items that are

measuring the same target construct, inferential knowledge about the character's sentiment and interaction.

More specifically, examinees were required to identify and evaluate the sentimental component of underlying interactions between the diverse characters in the text. Using four variations of item models with the products from topic modelling analysis, a total of 61,338 test items were generated from the seven books from the Harry Potter series. The validation of the item quality was conducted by evaluating the procedures and the products of each stage of the analysis. This evaluation was conducted to to ensure that the final product—the test items—are presenting the expected item format, item elements, and the item content and context. I addressed the summary of the findings and discussed the practical implications of the findings in the next sections.

Discussion of the Findings

Sentiment-weighted Latent Dirichlet Allocation Topic Modelling. The by-products of sentiment-topic modelling results had an integral role for generating inferential-type test items in my AIG framework. The underlying sentimental topics of the given text were modelled and represented using topic keywords and topic key sentences from the sentiment-weighted latent Dirichlet allocation approach. A total of 10 sentiment-related topics were identified from the Harry Potter series. All topic keywords belonged to the semantic categories of interest by representing the abstract social interaction and the psychological aspects, such as the sentiment or personality.

More specifically, a large proportion of the sentiment-topic keywords belonged to the semantic categories representing "an abstraction belonging to or characteristic of an entity". This semantic category could be further explained by its hyponyms, such as "human nature",

"cheerfulness", "character", and "personality". In addition, the sentiment-topic modelling results indicated that diverse topic dimensions underlying the stories could be captured and represented using topic keywords with an appropriate weighting scheme. This result was also evident with the drastic change of representative topic keywords when compared with the main topic modelling results, where no weighting scheme was applied.

The findings provide interesting implications and insights to educators and item developers in evaluating the quality and the characteristics of reading passages. The specific semantic category of topic keywords could be used to directly evaluate the appropriateness of the reading passages in generating quality inference-type questions. For instance, in this dissertation, six specific dimensions were identified and categorized: "group", "communication", "psychological features", "attributes", and "relation" (Table 18). This categorization was identified from one of the widely used lexical resources, WordNet.

The absence of robust and explicit guidelines in reading-passage evaluation can make it challenging to judge the quality of these types of test items thereby increasing the complexity of item development task. Traditionally, the appropriateness of reading passages for test item development was commonly determined by SMEs using the test blueprint. The blueprint works as a detailed guideline to generate test items based on the strands and the depth of knowledge targeted for different grade levels. For instance, the blueprint for inference item development in reading comprehension exam provides content-specific information as content-limit statements, such as "The item may ask the examinees to determine a theme or central ideas from a section of the passage or the entire passage (…)" or "The item may focus on the interaction of two or more story elements (…)" (Arizona Department of Education with American Institutes for Research, 2016). The guidelines aim to provide overarching evaluation criteria to identify which part of the

reading resources to focus on to generate test items. However, the overarching statement about the content and bias is often not a sufficient evaluation criteria to apply the standards in reading source evaluation. This is, in part, due to the complexity of drawing empricial evidence supporting the violation or the adherence of the contet to the strands in large-quantiy reading resources (Bråten, Braasch, & Salmerón, 2016).

Given such properties, test developers and item writers will be able to explicitly evaluate whether the given corpus provides an appropriate amount of sentimental components to construct test items within the validation process of the first stage. In other words, if the selected corpus provides a rich list of topic keywords within the six specific dimensions representing interactions and candidates, then the source text could be evaluated as "appropriate" and "sufficient" types of text. Moreover, for educators and item development, using the topic representation stemming from the specific semantic categories could effectively inform their teaching focus in reading comprehension. If the test items within such cognitive focus could create high-quality test items for operational uses, then teaching practices should also focus on identifying, extracting, and evaluating such properties from reading passages to make correct inferences in reading comprehension-focused instructions.

Rule-based Subtext Candidate Categorization and Evaluation. The rule-based subtext candidate categorization served three primary purposes in item generation. First, the candidate subtexts with no evident sentiment-topic structure were omitted to prevent item generation from unclear and uninformative reading passages as a source text. Second, the candidate subtexts could be parsed into two categories based on their sentiment-specific topic mixture distributions. Third, the processes of omitting inappropriate candidate subtexts and categorizing the subtexts

based on their topic distributions produced an important by-product, the sentiment-topic key sentences, which supported as the evidence of the underlying sentimental-topic structures.

The findings indicated that a total of 31,061 candidate subtexts could be identified as the two categories: coherent and divergent topic texts. This finding suggests that a significant number of subtexts could be parsed and created from a fictional story, in this case, the Harry Potter series. Then, the alignment between the final batch of candidate subtexts with their sentiment-topic keywords was evaluated. The evaluation revealed that the subtexts and its sentiment-topic keywords for each specific topic showed moderate to high (0.40 in average) correlation in terms of their overall sentiment (i.e., negative or positive) representation. This reveals new insights and empirical support to the integral role of the by-products of topic modelling (e.g., topic keywords and sentences) to explain the variability of sentimental representation in a text. Also, it highlights the plausibility of extracting supporting knowledge evidence from text using the modelling approaches, which resembles the previous cognitive modelling approach in a template-based automated item generation (e.g. Gierl & Lai, 2013; Gierl & Haladyna, 2012; Gierl, Lai, Hogan, & Matovinovic, 2015; Gierl, Lai & Matovinovic, 2020).

Test Item Generation with Four Item Models. The combination of four-item models from two categories with 31,061 candidate subtexts resulted in a total of 61,338 multiple choices (MC) items of more than four options. Overall, up to 3.35 variations of keyed options and 8.45 incorrect options or distractors could be generated. The items were could be categorized based on the types of the item models, which attempted to assess whether “the examineess could correctly identify the parts of the text (sentences) coherent with the given sentimental keyword” (coherent item model 1), “the examineess could identify the sentiment-topic keywords based on the highlighted parts of the text (sentences)” (coherent item model 2), “the examineess could

distinguish the varying sentiment between the different parts of the text and represent them as topic keywords" (divergent item model 1), and "the examineess could indicate parts of the text presenting the different sentimental topic" (divergent item model 2).

The item generation process was validated following the item quality guidelines by Ennis (1969, 1973, 1981), Collins et al. (1980), and Phillips (1989). The results indicated that my process satisfies the guidelines by addressing and implementing the integral item quality control components in the three primary stages of topic modelling, rule-based parsing, and the item modelling processes. An example in Figure 22 provided a step-by-step illustration to communicate how the unexpected outcomes (e.g., ambiguity in the item component) are controlled for to ensure that the final product follows the specified guidelines and expected outcomes.

The process of item generation with specific evidence highlighted (e.g., topic sentences as part of the text) and explained with their item generation rationale (e.g., the corresponding topic weigh) opens up an opportunity to objectively compare and evaluate the item quality from item writers and test developers. Because of the explicit and visible structure of the input and the output of the item models, the item writers and test developers will be able to comprehend how the items were constructed to evaluate specific domain knowledge. This reduces the inexplicit and black-box nature of manual test item writing and increases the replicability of the test item generation. The visible logical connection between the item elements (e.g., key sentences, topic keywords, and reading passages) would allow educators to prepare more effective feedback to examineess. For instance, examinees can be provided with specific examples about how the correct answers could be derived using the combinations of the key elements of items or, conversely, how the incorrect answers could be derived using the textual evidence. Using this

approach, the educators will be able to derive diagnostic evidence to understand the level of understanding of the examinees, based on their choice of an incorrect answer.

The visualization of the logical association between the reading sources and test items provides important validation evidence in reading comprehension item development. The strong alignment between the components in the reading sources and the test items has long been emphasized as an important soures of item quality in reading comprehension assessments (e.g., Katz, Lautenschlager, Blackbrun, & Harris, 1990; Royer, 1990; Freedle & Kostin, 1994; Kobayashi, 2002; Ozuru, Best, Bell, Witherspoon, & McNamara, 2007; Krumm, Hüffmeier, & Lievens, 2017). For instance, Katz et al. (1990) investigated the empirical validity of the test items extracted from the reading sections in the SAT. The findings indicated that the examineess did not encounter trouble answering test items correctly even when the reading passages were not presented. One of the findings revealed that the examineess could answer close to 70% of the test items correctly without referring to the source information passages.

This finding reinforces the importance of the strong alignment of evidence between the core elements in generating valid test items. This is particularly important in reading inference test items, in which the primary purpose of the assessment is to measure examineess' understanding of the reading passage as well as the activation of their background knowledge and experiences. Hence, the explicit visualization and the statistical evidence confirming the connections between the reading passage and the test item would serve as important validity evidence to generate high-quality test items in various test development settings.

To summarize, my AIG system using the Harry Potter series has demonstrated the capacity of the hybrid approach of item generation to produce a large number of test items in a previously challenging item type and domain. Three major stages of analysis were implemented

and validated to generate test items with the reading passages of divergent or coherent sentimental topics. Overall, the system demonstrated the capacity of generating a large number of test items from a fictional story with reduced guidance from SMEs. Moreover, the validation of each stage of item development potentially provides practical implications for item developers and educators. For instance, the appropriateness of the reading passages to generate inference-type questions could be explicitly evaluated. Also, more diagnostic and effective feedback could be prepared and provided to the item writers, test administrators, and students. The feedback for the item writers can be identified by comparing the topic representation and contribution (i.e., weights) or the distractors students chose with the correct answers. This way, item writers and test administrators will be able to effectively identify the level of understanding and misconceptions of the examinees for future test writing and instruction. The feedback for examinees could be provided by highlighting parts of the text which provide strong logical evidence supporting the correct, or incorrect, answers. By visually inspecting the highlighted parts of the text, examinees will be able to effectively reexamine their thought processes and learn effective skills for future tests. In essence, the hybrid approach of item generation provides the benefits of a data-informed approach with an educational framework that could increase the capacity of educational assessment.

Limitations and the Directions for Future Research

While the study was designed and conducted to minimize potential errors with the results and the interpretation, the following limitation should be carefully addressed for future research. The focus of this dissertation was to propose a hybrid approach of AIG, which extends the capacity of the previous two paradigms of item generation, template-based and non-template-based approaches. In this dissertation, I attempted to provide internal validation of the item

generation approach using theoretical and practical guidelines of item quality of each step (Ennis, 1969, 1973; 1981; Collins et al., 1980; Phillips, 1989). However, internal validation does not provide a complete understanding of the validity and the reliability of the generated test items. Validating test items is a fundamental issue in the field of AIG research, in which a large quantity, in our case 61,338 items, should be individually tested and evaluated. One possible approach to evaluate the content and the cognitive dimensions of the items are through reviewing a sample of test items using a substantive review. Second, the psychometric properties of the test items can be evaluated using a statistical review method. The following sections describe the importance of item validation with specific approaches to guide future research.

Substantive Methods for Evaluating the Test Item Quality. The substantive review of item quality focuses on assessing whether the generate test items meet the operational test administration standards and guidelines. This can be conducted by directly comparing the properties and the qualities of the generated test items with the ones written by SMEs. For a thorough review, SMEs should use a standardized rating scale to compare the quality of the two groups of test items. The example rating scale could include four choices indicating “the item is complete and requires no change”, “the item required minor revisions”, “the item requires major revisions from the item developer”, and “the item is flawed and should be rejected”.

To conduct this review in this dissertation framework, the SMEs should be provided with the same set of reading passages to generate inference-type test items. Then, the newly written test items will be compared with randomly sampled test items generated automatically using the current framework. The items will be reviewed by a committee of SMEs with extensive expertise to understand and evaluate the test items based on the item development standards. The evaluators should be blind to the item development process to avoid potential bias. The review

process will focus on rating each test item based on the standardized rating scale in terms of their completeness. Once the item evaluation is complete, the difference between the ratings of the two item groups will be evaluated. More specific guidelines about the substantive review are provided by Gierl, Latifi, Lai, Matovinovic, and Boughton (2016).

Substantive Methods for Evaluating the Item Models. The substantive review of the test item models could also provide important validation evidence of the item quality. This evaluation, in specific, could investigate the quality and the logic structure in item models to produce high-quality test items. The item models, in this dissertation, include four elements, which are the reading passages, stem, keyed-option, and distractors. To effectively evaluate the item model elements, Gierl and Lai (2016) recommended the evaluation using a validation table. A validation table should indicate the logical structure behind the extraction and generation of the keyed option and the distractors. Hence, the SMEs will be able to explicitly review and investigate how the information is connected and presented to the examineess to assess their knowledge. In this dissertation, the explicit evaluation of the logic should mostly focus on the association between the reading passages, topic sentences, and the topic keywords. The sentimental representation of the three elements should be displayed as a validation table and reviewed by SMEs to evaluate the quality of the item model and its logic structure. If the elements of the item models are correctly specified and placed, then the generated items should reflect the logical combinations of textual components required for making correct inferences. Hence, the substantive review of item models will ensure that the generated test items using the item models will include a correctly identified logical structure.

Statistical Methods for Evaluating Item Quality. Psychometric properties of the item (e.g., item difficulty and discrimination) using measurement models would be able to reveal

more insights about how the item behaves empirically. More specifically, the items could be evaluated based on how they discriminate examineess with different ability levels and how challenging it is to discern correct and incorrect answers for the examinees at certain ability levels. For instance, using a two-parameter model, or 2PL, in item response theory (IRT), the probability of correctly answering the items can be modelled using the difficulty parameter, b_i, discrimination parameter, a_i, and the ability of the examinee, θ_j (Equation 11). In this example, i denotes a test item and j indicates an examinee.

$$P_{ij}(\theta_j, b_i, a_i) = \frac{\exp\left[a_i(\theta_j - b_i)\right]}{1 + \exp[a_i(\theta_j - b_i)]} \tag{11}$$

A higher value of a_i would indicate that the item is more discriminating, thus providing more information about the examinee's ability succinctly. Typically in an operational administration setting, the discrimination parameter, a_i, commonly range from 0 to 0.4. Similarly, a higher value of b_i would indicate that the item is more difficult. To allow for such evaluation, it would be required to field test the generate items with a sufficient amount of sample sizes (de la Torre & Hong, 2010; de la Torre, Hong, & Deng, 2010). These psychometric properties would provide rich information about whether the items would behave successfully in an assessment empirically. While the evaluation of the item quality was out of the scope of this dissertation, future research should be conducted to collect and understand the statistical validity evidence of the automatically generated items.

References

ACT (2020). Preparing for the ACT Test. http://www.act.org/content/dam/act/unsecured/documents/Preparing-for-the-ACT.pdf

Agarwal, M. & Mannem, P. (2011, June) Automatic Gap-fill Question Generation from Textbooks. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), In Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, (pp. 56–64). Oregon: Association for Computational Linguistics.

Aist, G. (2001). Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. International Journal of Artificial Intelligence in Education, 12(2), 212-231.

Alves, C., Gierl, M., & Lai, H. (2010, April). Using Automated Item Generation to Promote Principled Test Design and Development. Paper presented at the American Educational Research Association. Denver, CO.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). Standards for educational and psychological testing (Rev.ed). Washington: American Educational Research Association.

Aquino, J. F., Chua, D. D., Kabiling, R. K., Pingco, J. N., & Sagum, R. (2011). Text2Test: Question generator utilizing information abstraction techniques and question generation methods for narrative and declarative text. In Proceedings of the 8th National Natural Language Processing Research Symposium (pp. 29-34).

Arizona Department of Education with American Institutes for Research (2016). ELA Item Specifications – Grade 7.

https://www.svjhscounseling.com/uploads/5/7/0/0/57006969/azmerit-ela-item-specs_grade-07.pdf

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. The Journal of Technology, Learning, and Assessment, 2(3). https://ejournals.bc.edu/index.php/jtla/article/view/1663

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3, 993-1022.

Bråten, I., Braasch, J. L., & Salmerón, L. (2016). Reading multiple and non-traditional texts: New opportunities and new challenges. In E. B. Moje, P. Afflerbach, P. Enciso, & N. K. Lesaux (Eds.), Handbook of reading research (Vol. V). New York, NY: Routledge.

Brown, J., Frishkoff, G., & Eskenazi, M. (2005, October). Automatic question generation for vocabulary assessment. In J. Tetreault, J. Burstein, & C. Leacock (Eds.) Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 819-826). Vancouver: Association for Computational Linguistics.

Chali, Y., & Hasan, S. A. (2015). Towards topic-to-question generation. Computational Linguistics, 41(1), 1-20.

Chatman, S. B. (1980). Story and discourse: Narrative structure in fiction and film. Cornell University Press.

Chen, C. Y., Liou, H. C., & Chang, J. S. (2006, July). FAST–An Automatic Generation System for Grammar Tests. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (pp. 1-4). Sydney: Association for Computational Linguistics.

Collins, A. M., Brown, J. S., & Larkin, K. M. (1980). Inferences in text understanding. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), Theoretical issues in reading comprehension (pp. 385 – 407). Hillsdale, NJ: Erlbaum.

Danon, G., & Last, M. (2017). A syntactic approach to domain-specific automatic question generation. arXiv preprint arXiv:1712.09827.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. Applied Psychological Measurement, 34(4), 267-285.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. Journal of Educational Measurement, 47(2), 227-249.

Debuse, J. C., & Lawley, M. (2016). Benefits and drawbacks of computer-based assessment and feedback systems: Student and educator perspectives. British Journal of Educational Technology, 47(2), 294-301.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. Applied Measurement in Education, 10(1), 61-82.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. Computer-based testing and the internet, 59-76.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. Educational measurement, 4, 471-515.

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106.

Ennis, R. H. (1969). Logic in teaching. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ennis, R. H. (1973). Inference. In H. S. Broudy, R. H. Ennis, & L. I. Krimerman (Eds.), Philosophy of educational research. New York: John Wiley & Sons, Inc.

Ennis, R. H. (1981). A conception of deductive logic competence. Teaching Philosophy, 4, 337-385.

Flor, M., & Riordan, B. (2018, June). A semantic role-based approach to open-domain automatic question generation. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.). Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications (pp. 254-263). New Orleans: Association for Computational Linguistics.

Freedle, R., & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? A reply to Katz, Lautenschlager, Blackburn, and Harris. Psychological Science, 5(2), 107-110.

Fuglede, B., & Topsoe, F. (2004, June). Jensen-Shannon divergence and Hilbert space embedding. Proceedings of the International Symposium on Information Theory (pp. 31). IEEE.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1), 177-196.

Gao, Y., Bing, L., Chen, W., Lyu, M. R., & King, I. (2018). Difficulty controllable generation of reading comprehension questions. arXiv preprint arXiv:1807.03586.

Gierl, M. J. and Lai, H. (2016). A process for reviewing and evaluating generated test items. Educational Measurement: Issues and Practice, 35, 6–20.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). Automatic item generation: Theory and practice. Routledge.

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. Educational Measurement: Issues and Practice, 32(3), 36-50.

Gierl, M. J., Bulut, O., & Zhang, X. (2018). Using computerized formative testing to support personalized learning in higher education: An application of two assessment technologies. In R. Zheng (Ed.), Digital technologies and instructional design for personalized learning (pp. 99-119). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-3940-7.ch005

Gierl, M. J., Lai, H., & Matovinovic, D. (2020). Augmented Intelligence and the Future of Item Development. Application of Artificial Intelligence to Assessment, 1.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. Medical Education, 46(8), 757–765

Gierl, M. J., Lai, H., Hogan, J. B. & Matovinovic, D. (2015). A method for generating educational test items that are aligned to the common core state standards. Journal for Applied Testing Technology, 16(1), 1-18.

Gierl, M. J., Lai, H., Houston, L., Rich, C., & Boughton, K. (2015). A Methodology for Generating Items in Three or More Languages Using Automated Processes. International Journal of e-Assessment, 1(1).

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. The Journal of Technology, Learning and Assessment, 7(2).

Gierl, M., Latifi, S. F., Lai, H., Matovinovic, D., & Boughton, K. A. (2016). Using Automated Procedures to Generate Test Items That Measure Junior High Science Achievement. In Handbook of Research on Technology Tools for Real-World Skill Development (pp. 590-610). IGI Global.

Griffths, T. L., and Steyvers, M. (2004), Finding Scientific Topics, Proceedings of the National Academy of Sciences of the United States of America, 101, 5228–5235.

Gütl, C., Lankmayr, K., Weinhofer, J., & Höfler, M. (2011). Enhanced Automatic Question Creator--EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. Electronic Journal of e-Learning, 9(1), 23-38.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. Journal of Second Language Writing, 3(1), 49-68. doi.org/10.1016/1060-3743(94)90005-1

Heilman, M. (2011). Automatic factual question generation from text. PhD thesis: Carnegie Mellon University.

Heilman, M., & Smith, N. A. (2009). Ranking automatically generated questions as a shared task. In The 2nd Workshop on Question Generation (Vol. 1, pp. 30-37).

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In F. Gey, M. Hearst, & R. Tong (Eds.). Proceedings of the 22nd annual international ACM SIGIR conference on

Research and development in information retrieval (pp. 50-57). New York: Association for Computing Machinery.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar & P. Resnick (Eds.), Proceedings of the eighth international AAAI conference on weblogs and social media (pp. 216–225). Palo Alto: AAAI Press.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. Psychological Science, 1(2), 122-127.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. Language testing, 19(2), 193-220.

Krumm, S., Hüffmeier, J., & Lievens, F. (2017). Experimental test validation: Examining the path € from test elements to test performance. European Journal of Psychological Assessment, 35, 225–232. https://doi.org/10.1027/1015-5759/a000393

Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466.

Kyllonen, P. C. (2009). New constructs, methods, and directions for computer-based assessment. The transition to computer-based assessment, 151-156.

Kyllonen, P. C., Pfeiffenberger. W, Trapani, C., & Weng, P. (2009). Evaluating Transfer Learning in College-Level Physics: Final Report. Princeton, NJ: Educational Testing Service.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2015). Handbook of test development. New York, NY: Routledge.

Liu, Y., Lv, N., Luo, J., & Yang, H. (2009, November). Subtopic Based Topic Evolution Analysis. In 2009 International Conference on Web Information Systems and Mining (pp. 168-172). IEEE.

Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying textbooks used in nursing education. Journal of Nursing Education, 40(1), 25-32.

Mazidi, K. (2017, April). Automatic question generation from passages. In International Conference on Computational Linguistics and Intelligent Text Processing (pp. 655-665). Springer, Cham.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Mitkov, R. & Ha, L.A. (2003, May). Computer-Aided Generation of Multiple-Choice Tests. In Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing (pp. 17 – 22). Association for Computational Linguistics.

Mitkov, R., Le An, H., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. Natural language engineering, 12(2), 177.

Monteiro Viera, J. M. & Brey, G. A. (2012). Automatic Topic Hierarchy Generation Using WordNet. In Digital Humanities.

Montoya, L. A., Egnatovitch, R., Eckhardt, E., Goldstein, M., Goldstein, R. A., & Steinberg, A. G. (2004). Translation challenges and strategies: The ASL translation of a computer-based, psychiatric diagnostic interview. Sign Language Studies, 4(4), 314-344.

Murtagh, F., Ganz, A., & McKie, S. (2009). The structure of narrative: the case of film scripts. Pattern Recognition, 42(2), 302-312.

Narayan, S., Simoes, G., Ma, J., Craighead, H., & Mcdonald, R. (2020). QURIOUS: Question Generation Pretraining for Text Generation. arXiv preprint arXiv:2004.11026.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 100-108). Association for Computational Linguistics.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. Cognition and Instruction, 25(4), 399-438.

Phillips, L. M. (1989). Developing and validating assessments of inference ability in reading comprehension, (Technical Report No 452). Champaign, IL: Centre for the Study of Reading, University of Illinois (ERIC Document Reproduction Service Number ED303 767).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Royer, J. (1990). The sentence verification technique: A new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), Cognitive assessment of language and math outcomes (pp. 144-191). Norwood, NJ: Ablex.

Rudner, L. (2010). Implementing the Graduate Management Admission Test computerized adaptive test. In W. van der Linden & C. Glas (Eds.), Elements of Adaptive Testing (p. 151-165), New York, NY: Springer.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC medical education, 16(1), 1-10.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. Educational measurement, 4, 307-353.

Singley, M., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), Item generation for test development (pp. 361-384). Mahwah, NJ: Lawrence Erlbaum.

Susanti, Y, Iida, R, Tokunaga, T (2015). Automatic generation of English vocabulary tests. In Proceedings of the 7th International Conference on Computer Supported Education. INSTICC, Lisbon, (pp. 77–87).

Susanti, Y., Tokunaga, T., & Nishikawa, H. (2020). Integrating automatic question generation with computerized adaptive testing. *Research and Practice in Technology Enhanced Learning, 15,* 1-22.

Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer-based assessment. Computers & Education, 56(4), 1032-1044.

The College Board. (2020). Official SAT Study Guide 2020 Edition. College Board.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009, June). Evaluation methods for topic models. In Proceedings of the 26th annual international conference on machine learning (pp. 1105-1112). Association for Computing Machinery.

Wilson, A., & Chew, P. A. (2010, June). Term weighting schemes for latent Dirichlet allocation. In human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics (pp. 465-473).

Wyse, B. & Piwek, P. (2009). Generating questions from Open Learn study units. In Craig, S. D. & Dicheva, D. (eds.), Proceedings of the 2nd Workshop on Question Generation, held at AIED 2009, pp. 66-73.

Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based question generation and implementation. Dialogue & Discourse, 3(2), 11-42.

Zhang, X. (2019). Using Automatic Item Generation to Create Content for Computerized Formative Assessment. Unpublished doctoral dissertation. University of Alberta.

Ziles, C., West, M., Herman, G. L., & Bretl, T. W. (2019). *Every University Should Have Computer-Based Testing Facility.* CSEDU 2019—Proceedings for the 11th International Conference on Computer Supported Eduation, 414-420.

Appendix A: Generated Items

A1: Items from Coherent Item Model 1

Q1. "Oh, shut up," said Hermione, but she agreed to go and watch out for Snape. "And we'd better stay outside the third-floor corridor," Harry told Ron. "Come on. " No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper. "I suppose you think you're harder to get past than a pack of enchantments!" "Enough of this nonsense! If I hear you've come anywhere near here again, I'll take another fifty points from Gryffindor! Yes, Weasley, from my own House!" Harry and Ron went back to the common room. Harry had just said, "At least Hermione's on Snape's tail," when the portrait of the Fat Lady swung open and Hermione came in. "I'm sorry, Harry!" "Snape came out and asked me what I was doing, so I said I was waiting for Flitwick, and Snape went to get him, and I've only just got away, I don't know where Snape went. " "Well, that's it then, isn't it?" Harry said. "I'm going out of here tonight and I'm going to try and get to the Stone first. " "You're mad!" said Ron. "You can't!" said Hermione. "After what McGonagall and Snape have said? You'll be expelled!" "SO WHAT?" Harry shouted. "Don't you understand? If Snape gets hold of the Stone, Voldemort's coming back! There won't be any Hogwarts to get expelled from! D'you think he'll leave you and your families alone if Gryffindor wins the House Cup? If I get caught before I can get to the Stone, well, I'll have to go back to the Dursleys and wait for Voldemort to find me there, it's only dying a bit later than I would have, because I'm never going over to the Dark Side! Voldemort killed my parents, remember?" "You're right, Harry," said Hermione in a small voice. "I'll use the Invisibility Cloak," said Harry. "It's just lucky I got it back. " "But will it cover all three of us?" said Ron. "All — all three of us?" "Oh, come off it, you don't think we'd let you go alone?" "Of course not," said Hermione briskly.

The main characters' (Harry, Hermione, Ron) feeling "**scared**" is most likely related to the statement:

a. **"No sooner had they reached the door separating Fluffy from the rest of the school than Professor McGonagall turned up again and this time, she lost her temper."***
b. Harry and Ron went back to the common room. Harry had just said, "At least Hermione's on Snapes' tail."
c. "Oh, come off it, you don't think we'd let you go alone?" "Of course not" said Hermione briskly.
d. "I'm never going over to the Dark side! Voldemort killed my parents, remember?"

Q2. Harry saw a familiar, snow-white marble building in the distance — Gringotts Bank. Hagrid had steered him right into Diagon Alley. "Yer a mess!" said Hagrid gruffly, brushing soot off Harry so forcefully he nearly knocked him into a barrel of dragon dung outside an apothecary. "Skulkin' around Knockturn Alley, I dunno — dodgy place, Harry — don' want no one ter see yeh down there — " "I realized that," said Harry, ducking as Hagrid made to brush him off again. "I told you, I was lost — what were you doing down there, anyway?" " I was lookin' fer a Flesh-Eatin' Slug Repellent," growled Hagrid. "They're ruinin' the school cabbages. Yer not on yer own?" "I'm staying with the Weasleys but we got separated," Harry explained. "I've got to go and find them. . " They set off together down the street. "How come yeh never wrote back ter me?" said Hagrid as Harry jogged alongside him (he had to take three steps to every stride of Hagrid 's enormous boots). Harry explained all about Dobby and the Dursleys. "Lousy Muggles," growled Hagrid. "If I'd've known — " "Harry! Harry! Over here!" Harry looked up and saw Hermione Granger standing at the top of the white flight of steps to Gringotts. "What happened to your glasses? Hello, Hagrid — Oh, it's wonderful to see you two again — Are you coming into Gringotts, Harry?" "As soon as I've found the Weasleys," said Harry. "Yeh won't have long ter wait," Hagrid said with a grin. Harry and Hermione looked around: Sprinting up the crowded street were Ron, Fred, George, Percy, and Mister Weasley. "Harry," Mister Weasley panted. "We hoped you'd only gone one grate too far. " He mopped his glistening bald patch. "Molly's frantic — she's coming now — " "Where did you come out?" Ron asked. "Knockturn Alley," said Hagrid grimly. "Excellent. " said Fred and George together. "We've never been allowed in," said Ron enviously. "I should ruddy well think not," growled Hagrid.

The main character's (Harry) feeling "welcomed" is most likely related to the statement:

a. "Hello, Hagrid – Oh, it's wonderful to see you two again – Are you coming into Gringotts, Harry?"*
b. "I should ruddy well think not," growled Hagrid.
c. "Skulkin' around Knockturn Alley, I dunno — dodgy place, Harry — don' want no one ter see yeh down there — "
d. "I'm staying with the Weasleys but we got separated," Harry explained.

Q3. "Yeah, it does!" said Ron excitedly, but Sirius shook his head. "Listen, if Crouch wants to investigate Snape, why hasn't he been coming to judge the tournament? It would be an ideal excuse to make regular visits to Hogwarts and keep an eye on him. " "So you think Snape could be up to something, then?" asked Harry, but Hermione broke in. "Look, I don't care what you say, Dumbledore trusts Snape — " "Oh give it a rest, Hermione," said Ron impatiently. "I know Dumbledore 's brilliant and everything, but that doesn't mean a really clever Dark wizard couldn't fool him — " "Why did Snape save Harry's life in the first year, then? Why didn't he just let him die?" "I dunno — maybe he thought Dumbledore would kick him out — " "What d'you think, Sirius?" Harry said loudly, and Ron and Hermione stopped bickering to listen. "I think they've

both got a point," said Sirius, looking thoughtfully at Ron and Hermione. "Ever since I found out Snape was teaching here, I've wondered why Dumbledore hired him. Snape 's always been fascinated by the Dark Arts, he was famous for it at school. Slimy, oily, greasy-haired kid, he was," Sirius added, and Harry and Ron grinned at each other. "Snape knew more curses when he arrived at school than half the kids in seventh year, and he was part of a gang of Slytherins who nearly all turned out to be Death Eaters. " Sirius held up his fingers and began ticking off names. "Rosier and Wilkes — they were both killed by Aurors the year before Voldemort fell. The Lestranges — they're a married couple — they're in Azkaban. But as far as I know, Snape was never even accused of being a Death Eater — not that that means much. And Snape 's certainly clever and cunning enough to keep himself out of trouble. " "Snape knows Karkaroff pretty well, but he wants to keep that quiet," said Ron. "Yeah, you should've seen Snape's face when Karkaroff turned up in Potions yesterday!" said Harry quickly. "Karkaroff wanted to talk to Snape, he says Snape's been avoiding him. Karkaroff looked really worried.

The main character's (Harry) feeling "fooled" is most likely related to the statement:

a. **"Why did he fake his death?" "Because he knew you were about to kill him like you killed my parents!"***
b. "He approached Lupin and the struggling rat, and his wet eyes suddenly seemed to be burning in his face."
c. "I persuaded Lily and James to change to Peter at the last moment, persuaded them to use him as Secret- Keeper instead of me"
d. "If he really is a rat, it won't hurt him."

Q4. Ron seemed very sleepy too, though Harry could not see why he should be. Harry's third detention passed in the same way as the previous two, except that after two hours the words "I must not tell lies" did not fade from the back of Harry's hand, but remained scratched there, oozing droplets of blood. The pause in the pointed quill's scratching made Professor Umbridge look up. "Ah," she said softly, moving around her desk to examine his hand herself. "Good. You may leave for tonight. " "Do I still have to come back tomorrow?" said Harry, picking up his schoolbag with his left hand rather than his smarting right. "Oh yes," said Professor Umbridge, smiling widely as before. "Yes, I think we can etch the message a little deeper with another evening's work. " He had never before considered the possibility that there might be another teacher in the world he hated more than Snape, but as he walked back toward Gryffindor Tower he had to admit he had found a contender. She's evil, he thought, as he climbed a staircase to the seventh floor, she's an evil, twisted, mad, old — "Ron?" He had reached the top of the stairs, turned right, and almost walked into Ron, who was lurking behind a statue of Lachlan the Lanky, clutching his broomstick. He gave a great leap of surprise when he saw Harry and attempted to hide his new Cleansweep Eleven behind his back. "What are you doing?" "Er — nothing. What are you doing?" Harry frowned at him. "Come on, you can tell me! What are you hiding here for?" "I'm — I'm hiding from Fred and

George, if you must know," said Ron. "They just went past with a bunch of first years, I bet they're testing stuff on them again, I mean, they can't do it in the common room now, can they, not with Hermione there. " "But what have you got your broom for, you haven't been flying, have you?" Harry asked. "I — well — well, okay, I'll tell you, but don't laugh, all right?" Ron said defensively, turning redder with every second.

The main character's (Harry) feeling "furious" is most likely related to the statement:

a. "**He had never before considered the possibility that there might be another teacher in the world but as he walked back toward Gryffindor Tower he had to admit he had found a contender."**
b. **"She's evil, he thought, as he climbed a staircase to the seventh floor, she's an evil, twisted, mad, old"**
c. "Harry's third detention passed in the same way as the previous two, except that after two hours the words"
d. "I must not tell lies" did not fade from the back of Harry's hand, but remained scratched there, oozing droplets of blood.

Q5. Harry protested. Cho shouted, " Cedric gave me loads of Chocolate Frog cards, look!" And she pulled out fistfuls of cards from inside her robes and threw them into the air, and then turned into Hermione, who said, "You did promise her, you know, Harry. I think you'd better give her something else instead. How about your Firebolt?" And Harry was protesting that he could not give Cho his Firebolt because Umbridge had it, and anyway the whole thing was ridiculous, he'd only come to the D. A. room to put up some Christmas baubles shaped like Dobby's head. Harry put out his tongue. Harry longed to bite the man . But the man was stirring . a silvery cloak fell from his legs as he jumped to his feet; and Harry saw his vibrant, blurred outline towering above him, saw a wand withdrawn from a belt. It was aching fit to burst. "Harry! HARRY!" "Harryl" Ron was standing over him looking extremely frightened. There were more figures at the foot of Harry's bed. "He's really ill," said a scared voice. "Should we call someone?" "Harry! Harryl" He had to tell Ron, it was very important that he tell him. Taking great gulps of air, Harry pushed himself up in bed, willing himself not to throw up again, the pain half-blinding him. "Your dad," he panted, his chest heaving. "Your dad's been attacked. " "What?" said Ron uncomprehendingly. "Your dad! He's been bitten, it's serious, there was blood everywhere. " "I'm going for help," said the same scared voice, and Harry heard footsteps running out of the dormitory. "Harry, mate," said Ron uncertainly, "you . you were just dreaming. " "No!" said Harry furiously; it was crucial that Ron understand. "It wasn't a dream . not an ordinary dream. I did it. " He could hear Seamus and Dean muttering but did not care. He retched again and Ron leapt backward out of the way. "Harry, you're not well," he said shakily. "Neville's gone for help. " "I'm fine!"

The main character's (Ron) feeling "terrified" is most likely related to the statement:

a. “**Harry! Ron was standing over him looking extremely frightened. “He's really ill,” said a scared voice.”***
b. **“**You did promise her, you know, Harry. I think you'd better give her something else instead. How about your Firebolt? ”
c. “Your dad! He's been bitten, it's serious, there was blood everywhere.”
d. **“**It wasn't a dream. not an ordinary dream. I did it. " He could hear Seamus and Dean muttering but did not care.”

Q6. And then his scar burst open and he was Voldemort and he was running across the fetid bedroom, his long white hands clutching at the windowsill as he glimpsed the bald man and the little woman twist and vanish, and he screamed with rage, a scream that mingled with the girl's, that echoed across the dark gardens over the church bells ringing in Christmas Day. And his scream was Harry's scream, his pain was Harry's pain that it could happen here, where it had happened before . here, within sight of that house where he had come so close to knowing what it was to die. to die. "Nice costume, mister!" The gate creaked a little as he pushed it open, but James Potter did not hear. He was over the threshold as James came sprinting into the hall. . "Lily, take Harry and go! Ill hold him off!" "Avada Kedavra!" The green light filled the cramped hallway, it lit the pram pushed against the wall, it made the banisters glare like lightning rods, and James Potter fell like a marionette whose strings were cut. "Not Harry, not Harry, please not Harry!" "Stand aside, you silly girl . stand aside, now. " "Not Harry, please no, take me, kill me instead — " "This is my last warning — " "Not Harry! Not Harry! Not Harry! Please — I'll do anything — " "Stand aside. Stand aside, girl!" The child began to cry: It had seen that he was not James. He did not like it crying, he had never been able to stomach the small ones whining in the orphanage — "Avada Kedavra!" "No," he moaned. "No . " And now he stood at the broken window of Bathilda's house, immersed in memories of his greatest loss, and at his feet the great snake slithered over broken china and glass. "No . " "Harry, it's all right, you're all right!" "No. " "Harry, it's okay, wake up, wake up!" He was Harry. . Harry, not Voldemort .

The main charater's (Harry’s family) feeling "painful" is most likely related to the statement

a. “And his scream was Harry's scream, his pain was Harry's pain .",
b. “And now he stood at the broken window of Bathilda's house, immersed in memories of his greatest loss, and at his feet the great snake slithered over broken china and glass.”
c. “He did not like it crying, he had never been able to stomach the small ones whining in the orphanage — "
d. **"Not Harry, please no, take me, kill me instead — "***

A2: Items from Coherent Item Model 2

Q7. Could all this have anything to do with the Potters? The Dursleys got into bed. Miss Dursley fell asleep quickly, but Mister Dursley lay awake, turning it all over in his mind. His last, comforting thought before he fell asleep was that even if the Potters were involved, there was no reason for them to come near him and Miss Dursley. The Potters knew very well what he and Petunia thought about them and their kind. He couldn't see how he and Petunia could get mixed up in anything that might be going on — he yawned and turned over — it couldn't affect them. Mister Dursley might have been drifting into an uneasy sleep, but the cat on the wall outside was showing no sign of sleepiness. This man's name was Albus Dumbledore. Albus Dumbledore didn't seem to realize that he had just arrived in a street where everything from his name to his boots was unwelcome. He chuckled and muttered, "I should have known. " Dumbledore slipped the Put- Outer back inside his cloak and set off down the street toward number four, where he sat down on the wall next to the cat. "Fancy seeing you here, Professor McGonagall. " "How did you know it was me?" "My dear Professor, I've never seen a cat sit so stiffly. " "You'd be stiff if you'd been sitting on a brick wall all day," said Professor McGonagall. "All day? I must have passed a dozen feasts and parties on my way here. " Professor McGonagall sniffed angrily. "Oh yes, everyone's celebrating, all right," she said impatiently. "You'd think they'd be a bit more careful, but no — even the Muggles have noticed something's going on. It was on their news. " She jerked her head back at the Dursleys' dark living-room window. "I heard it. Shooting stars down in Kent — I'll bet that was Dedalus Diggle. He never had much sense. " "You can't blame them," said Dumbledore gently. "We've had precious little to celebrate for eleven years. " "I know that," said Professor McGonagall irritably. "But that's no reason to lose our heads.

What can be reasonably inferred from line 19 of the passage **“** "I know that," said Professor McGonagall irritably.'"But that's no reason to lose our heads.**”** that the character (Professor McGonagall) felt:

a. Annoyed	d. Frightened	g. Guilty
b. Worried	e. Terrified	h. Hatred
c. Unpleasant	f. Struggled	

Q8. He had had it as long as he could remember, and the first question he could ever remember asking his Aunt Petunia was how he had gotten it. "In the car crash when your parents died," she had said. "And don't ask questions. " Don't ask questions — that was the first rule for a quiet life with the Dursleys. Uncle Vernon entered the kitchen as Harry was turning over the bacon. "Comb your hair!" About once a week, Uncle Vernon looked over the top of his newspaper and shouted that Harry needed a haircut. Harry must have had more haircuts than the rest of the boys in his class put together, but it made no difference, his hair simply grew that way — all over the place. Harry was frying eggs by the time Dudley arrived in the kitchen with his mother. Dudley looked a lot like Uncle Vernon. Aunt Petunia often said that Dudley looked like a baby angel —

Harry often said that Dudley looked like a pig in a wig. Harry put the plates of egg and bacon on the table, which was difficult as there wasn't much room. Dudley, meanwhile, was counting his presents. "Thirty-six," he said, looking up at his mother and father. "That's two less than last year. " "Darling, you haven't counted Auntie Marge's present, see, it's here under this big one from Mommy and Daddy. " "All right, thirty-seven then," said Dudley, going red in the face. Harry, who could see a huge Dudley tantrum coming on, began wolfing down his bacon as fast as possible in case Dudley turned the table over. Aunt Petunia obviously scented danger, too, because she said quickly, "And we'll buy you another two presents while we're out today. Is that all right?" Dudley thought for a moment. Finally he said slowly, "So I'll have thirty . thirty . " "Thirty-nine, sweetums," said Aunt Petunia. "Oh. " Dudley sat down heavily and grabbed the nearest parcel. "All right then. " Uncle Vernon chuckled. "Little tyke wants his money's worth, just like his father. 'Atta boy, Dudley!" He ruffled Dudley's hair.

What can be reasonably inferred from line 2-3 of the passage " "In the car crash when your parents died," she had said. " And don't ask questions." " that the main character (Harry) felt:

a. Curious	d. Terrified	g. Danger
b. Ignored	e. Rage	h. Scared
c. Discouraged	f. Temper	

What can be reasonably inferred from line 16-17 of the passage "Aunt Petunia obviously scented danger, too, because she said quickly, "And we'll buy you another two presents while we're out today. Is that all right?" " that that the character (Aunt Petunia) felt:

a. Alarmed	d. Scared	g. Generous
b. Worried	e. Danger	h. Hopeful
c. Terrified	f. Relieved	

Q9. A chink of sky was visible between the heavy curtains: It was the cool, clear blue of watered ink, somewhere between night and dawn, and everything was quiet except for Ron and Hermione's slow, deep breathing. Harry glanced over at the dark shapes they made on the floor beside him. Ron had had a fit of gallantry and insisted that Hermione sleep on the cushions from the sofa, so that her silhouette was raised above his. Her arm curved to the floor, her fingers inches from Ron's. Harry wondered whether they had fallen asleep holding hands. He lay on the floor and he thought of the Horcruxes, of the daunting, complex mission Dumbledore had left him. . Dumbledore . The grief that had possessed him since Dumbledore's death felt different now. The accusations he had heard from Muriel at the wedding seemed to have nested in his brain like diseased things, infecting his memories of the wizard he had idolized. Could Dumbledore have let such things happen? Had he been like Dudley, content to watch neglect and abuse as long as it did not affect him? Harry thought of Godric's Hollow, of graves Dumbledore

had never mentioned there; he thought of mysterious objects left without explanation in Dumbledore's will, and resentment swelled in the darkness. Why hadn't Dumbledore told him? Had Dumbledore actually cared about Harry at all? Or had Harry been nothing more than a tool to be polished and honed, but not trusted, never confided in? Harry could not stand lying there with nothing but bitter thoughts for company. On the landing he whispered, "Lumos," and started to climb the stairs by wandlight. On the second landing was the bedroom in which he and Ron had slept last time they had been here; he glanced into it. Harry remembered the overturned troll leg downstairs. Snape? Or perhaps Mundungus, who had pilfered plenty from this house both before and after Sirius died? Harry's gaze wandered to the portrait that sometimes contained Phineas Nigellus Black, Sirius's great-great-grandfather, but it was empty, showing nothing but a stretch of muddy backdrop.

What can be reasonably inferred from line 14-15 of the passage " "Had Dumbledore actually cared about Harry at all? Or had Harry been nothing more than a tool to be polished and honed, but not trusted, never confided in?" " that that the main character (Harry) felt:

a. Furious	d. Frightened	g. Excited
b. Rage	e. Loved	h. Attacked
c. Hatred	f. Destoryed	

Q10. "C'mon, Harry, you've never missed the Snitch before. " "There had to be one time you didn't get it," said George. "It's not over yet," said Fred. "We lost by a hundred points, right? So if Hufflepuff loses to Ravenclaw and we beat Ravenclaw and Slytherin . " "Hufflepuff'll have to lose by at least two hundred points," said George. "But if they beat Ravenclaw . " "No way, Ravenclaw is too good. But if Slytherin loses against Hufflepuff . " "It all depends on the points — a margin of a hundred either way — " Harry lay there, not saying a word. They had lost . for the first time ever, he had lost a Quidditch match. After ten minutes or so, Madam Pomfrey came over to tell the team to leave him in peace. "Well come and see you later," Fred told him. "Don't beat yourself up, Harry, you're still the best Seeker we've ever had. " Madam Pomfrey shut the door behind them, looking disapproving. Ron and Hermione moved nearer to Harry's bed. "Dumbledore was really angry," Hermione said in a quaking voice. "I've never seen him like that before. We heard him — " "Then he magicked you onto a stretcher," said Ron. "And walked up to school with you floating on it. Everyone thought you were . " His voice faded, but Harry hardly noticed. He looked up and saw Ron and Hermione looking at him so anxiously that he quickly cast around for something matter-of-fact to say. "Did someone get my Nimbus?" Ron and Hermione looked quickly at each other. "Er — " "What?" said Harry, looking from one to the other. "Well . when you fell off, it got blown away," said Hermione hesitantly. "And?" "And it hit — it hit — oh, Harry — it hit the Whomping Willow. " Harry's insides lurched. "And?" "Well, you know the Whomping Willow," said Ron. "It — it doesn't like being hit. " "Professor Flitwick brought it back just before you came around," said Hermione in a very small voice.

What can be reasonably inferred from line 3-4 of the passage "Hufflepuff'll have to lose by at least two hundred points," said George. " that the character (Geroge) felt:

a. Hopeless	d. Lucky	g. Delighted
b. Doubtful	e. Excited	h. Scared
c. Losing	f. Clever	

Q11. "Who said anythin' abou' giants? Who's told yeh what I've — who's said I've bin — eh?" "We guessed," said Hermione apologetically. "Oh, yeh did, did yeh?" said Hagrid, fixing her sternly with the eye that was not hidden by the steak. "It was kind of . obvious," said Ron. Harry nodded. Hagrid glared at them, then snorted, threw the steak onto the table again and strode back to the kettle, which was now whistling. "Never known kids like you three fer knowin' more'n yeh oughta," he muttered, splashing boiling water into three of his bucket-shaped mugs. "An' I'm not complimentin' yeh, neither. Interferin'. " "So you have been to look for giants?" said Harry, grinning as he sat down at the table. Hagrid set tea in front of each of them, sat down, picked up his steak again, and slapped it back over his face. "Yeah, all righ'," he grunted, "I have. " "And you found them?" said Hermione in a hushed voice. "Well, they're not that difficult ter find, ter be honest," said Hagrid. "Pretty big, see. " "Where are they?" said Ron. "Mountains," said Hagrid unhelpfully. "So why don't Muggles — ?" "They do," said Hagrid darkly. "O'ny their deaths are always put down ter mountaineerin' accidents, aren' they?" "Come on, Hagrid, tell us what you've been up to!" said Ron. "Tell us about being attacked by the giants and Harry can tell you about being attacked by the dementors — " Hagrid choked in his mug and dropped his steak at the same time; a large quantity of spit, tea, and dragon blood was sprayed over the table as Hagrid coughed and spluttered and the steak slid, with a soft splat, onto the floor. "Whadda yeh mean, attacked by dementors?" growled Hagrid. "Didn't you know?" Hermione asked him, wide-eyed. "I don' know anything that's been happenin' since I left. Yeh're not serious?" "Yeah, I am, they turned up in Little Whinging and attacked my cousin and me, and then the Ministry of Magic expelled me — " "WHAT?" " — and I had to go to a hearing and everything, but tell us about the giants first. " "You were expelled?" "Tell us about your summer and I'll tell you about mine. " Hagrid glared at him through his one open eye. Harry looked right back, an expression of innocent determination on his face. "Oh, all righ'," Hagrid said in a resigned voice. He bent down and tugged the dragon steak out of Fang's mouth. "Oh, Hagrid, don't, it's not hygien — " Hermione began, but Hagrid had already slapped the meat back over his swollen eye. He took another fortifying gulp of tea and then said, "Well, we set off righ' after term ended — " "Madame Maxime went with you, then?" Hermione interjected.

What can be reasonably inferred from line 13-16 of the passage "Tell us about being attacked by the giants and Harry can tell you about being attacked by the dementors — " Hagrid choked in his mug and dropped his steak at the same time; a large quantity of spit, tea, and dragon blood

was sprayed over the table as Hagrid coughed and spluttered and the steak slid, with a soft splat, onto the floor." That the character (Hagrid) felt:

a. Horrified	d. Disappointed	g. Awful
b. Impatient	e. Petrified	h. Excited
c. Surprised	f. Mad	

Q12. "What?" "You're supposed to be learning how to close your mind to this sort of thing," said Hermione, suddenly stern. "I know I am," said Harry. "But — " "Well, I think we should just try and forget what you saw," said Hermione firmly. "And you ought to put in a bit more effort on your Occlumency from now on. " Harry was so angry with her that he did not talk to her for the rest of the day, which proved to be another bad one. When people were not discussing the escaped Death Eaters in the corridors today, they were laughing at Gryffindor's abysmal performance in their match against Hufflepuff; the Slytherins were singing "Weasley Is Our King" so loudly and frequently that by sundown Filch had banned it from the corridors out of sheer irritation. The week did not improve as it progressed: Harry received two more D's in Potions, was still on tenterhooks that Hagrid might get the sack, and could not stop himself from dwelling on the dream in which he had seen Voldemort, though he did not bring it up with Ron and Hermione again because he did not want another telling-off from Hermione. He wished very much that he could have talked to Sirius about it, but that was out of the question, so he tried to push the matter to the back of his mind. "Get up, Potter. " A couple of weeks after his dream of Rookwood, Harry was to be found, yet again, kneeling on the floor of Snape's office, trying to clear his head. He had just been forced, yet again, to relive a stream of very early memories he had not even realized he still had, most of them concerning humiliations Dudley and his gang had inflicted upon him in primary school. "That last memory," said Snape. "What was it?" "I don't know," said Harry, getting wearily to his feet. He was finding it increasingly difficult to disentangle separate memories from the rush of images and sound that Snape kept calling forth. "You mean the one where my cousin tried to make me stand in the toilet?" "No," said Snape softly. "I mean the one concerning a man kneeling in the middle of a darkened room. " "It's . nothing," said Harry. Snape's dark eyes bored into Harry's. Remembering what Snape had said about eye contact being crucial to Legilimency, Harry blinked and looked away. "How do that man and that room come to be inside your head, Potter?" said Snape. "It — " said Harry, looking everywhere but at Snape, "it was — just a dream I had. " "A dream," repeated Snape. There was a pause during which Harry stared fixedly at a large dead frog suspended in a purple liquid in its jar. "You do know why we are here, don't you, Potter?" said Snape in a low, dangerous voice. What can be reasonably inferred from line 24-25 of the passage "There was a pause during which Harry stared fixedly at a large dead frog suspended in a purple liquid in its jar." That the character (Harry) felt:

a. Forbidden	d. Impatient	g. Lost
b. Frightened	e. Feared	h. Drowned
c. Trapped	f. Miserable	

Q13. Harry too had recognized the slow-bubbling, mudlike substance in the second cauldron, but did not resent Hermione getting the credit for answering the question; she, after all, was the one who had succeeded in making it, back in their second year. "Excellent, excellent! Now, this one here . yes, my dear?" said Slughorn, now looking slightly bemused, as Hermione's hand punched the air again. "It's Amortentia!" "It is indeed. It seems almost foolish to ask," said Slughorn, who was looking mightily impressed, "but I assume you know what it does?" "It's the most powerful love potion in the world!" said Hermione. "Quite right! You recognized it, I suppose, by its distinctive mother-of-pearl sheen?" "And the steam rising in characteristic spirals," said Hermione enthusiastically, "and it's supposed to smell differently to each of us, according to what attracts us, and I can smell freshly mown grass and new parchment and — " But she turned slightly pink and did not complete the sentence. "May I ask your name, my dear?" said Slughorn, ignoring Hermione's embarrassment. "Hermione Granger, sir. " "Granger? Can you possibly be related to Hector Dagworth-Granger, who founded the Most Extraordinary Society of Potioneers?" "No, I don't think so, sir. I'm Muggle-born, you see. " Harry saw Malfoy lean close to Nott and whisper something; both of them sniggered, but Slughorn showed no dismay; on the contrary, he beamed and looked from Hermione to Harry, who was sitting next to her. "Oho! ‘One of my best friends is Muggle-born, and she's the best in our yeah' I'm assuming this is **the** very friend of whom you spoke, Harry?" "Yes, sir," said Harry. "Well, well, take twenty well-earned points for Gryffindor, Miss Granger," said Slughorn genially. Malfoy looked rather as he had done the time Hermione had punched him in the face. Hermione turned to Harry with a radiant expression and whispered, "Did you really tell him I'm the best in the year? Oh, Harry!" "Well, what's so impressive about that?" whispered Ron, who for some reason looked annoyed. "You are the best in the year — I'd've told him so if he'd asked me!" Hermione smiled but made a "shhing" gesture, so that they could hear what Slughorn was saying. Ron looked slightly disgruntled. "Amortentia doesn't really create love, of course. It is probably the most dangerous and powerful potion in this room — oh yes," he said, nodding gravely at Malfoy and Nott, both of whom were smirking skeptically. "When you have seen as much of life as I have, you will not underestimate the power of obsessive love. . "And now," said Slughorn, "it is time for us to start work. " "Sir, you haven't told us what's in this one," said Ernie Macmillan, pointing at a small black cauldron standing on Slughorn 's desk. "Oho," said Slughorn again. Harry was sure that Slughorn had not forgotten the potion at all, but had waited to be asked for dramatic effect.

What can be reasonably inferred from line 6-8 of the passage “It's the most powerful love potion in the world!" said Hermione. "Quite right! You recognized it, I suppose, by its distinctive mother-of-pearl sheen?" That the character (Hermione) felt:

a. Excited	d. Impatient	g. Punished
b. Enjoying	e. Comfortable	h. Rushed
c. Interested	f. Powerful	

A3: Items from Divergent Item Model 1

Q14. They didn't think they could bear it if anyone found out about the Potters. Miss Potter was Miss Dursley's sister, but they hadn't met for several years; in fact, Miss Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that. Mister Dursley hummed as he picked out his most boring tie for work, and Miss Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair. (A) At half past eight, Mister Dursley picked up his briefcase, pecked Miss Dursley on the cheek, and tried to kiss Dudley good-bye but missed, because Dudley was now having a tantrum and throwing his cereal at the walls. "Little tyke," chortled Mister Dursley as he left the house. For a second, Mister Dursley didn't realize what he had seen — then he jerked his head around to look again. (B) Mister Dursley blinked and stared at the cat. As Mister Dursley drove around the corner and up the road, he watched the cat in his mirror. Mister Dursley gave himself a little shake and put the cat out of his mind. Mister Dursley couldn't bear people who dressed in funny clothes — the getups you saw on young people! Mister Dursley was enraged to see that a couple of them weren't young at all; why, that man had to be older than he was, and wearing an emerald-green cloak! But then it struck Mister Dursley that this was probably some silly stunt — these people were obviously collecting for something . yes, that would be it. The traffic moved on and a few minutes later, Mister Dursley arrived in the Grunnings parking lot, his mind back on drills. Mister Dursley always sat with his back to the window in his office on the ninth floor. Mister Dursley, however, had a perfectly normal, owl-free morning.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “delighted” in (A) while the character felt “confused” in (B).
b. The character felt “amused” in (A) while the character felt “cheerful” in (B).
c. The character felt “lucky“ in (A) while the character felt “hopeful” in (B).
d. The character felt “treated” in (A) while the character felt “suspicious” in (B).

Q15. Before he could think of what to say, however, Sirius had beckoned him to his side. (A)"I want you to take this," he said quietly, thrusting a badly wrapped package roughly the size of a paperback book into Harry's hands. "What is it?" Harry asked. "A way of letting me know if Snape's giving you a hard time. No, don't open it in here!" said Sirius, with a wary look at Miss Weasley, who was trying to persuade the twins to wear hand-knitted mittens. "I doubt Molly would approve — but I want you to use it if you need me, all right?" "Okay," said Harry, stowing the package away in the inside pocket of his jacket, but he knew he would never use whatever it was.(B) It would not be he, Harry, who lured Sirius from his place of safety, no matter how

foully Snape treated him in their forthcoming Occlumency classes. "Let's go, then," said Sirius, clapping Harry on the shoulder and smiling grimly, and before Harry could say anything else, they were heading upstairs, stopping before the heavily chained and bolted front door, surrounded by Weasleys. "Good-bye, Harry, take care," said Miss Weasley, hugging him. "See you Harry, and keep an eye out for snakes for me!" said Mister Weasley genially, shaking his hand. "Right — yeah," said Harry distractedly. It was his last chance to tell Sirius to be careful; he turned, looked into his godfather's face and opened his mouth to speak, but before he could do so Sirius was giving him a brief, one-armed hug. He said gruffly, "Look after yourself, Harry," and next moment Harry found himself being shunted out into the icy winter air, with Tonks (today heavily disguised as a tall, tweedy woman with iron-gray hair) chivvying him down the steps. They followed Lupin down the front steps. As he reached the pavement, Harry looked around. "Come on, the quicker we get on the bus the better," said Tonks, and Harry thought there was nervousness in the glance she threw around the square. Lupin flung out his right arm.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “trusting” in (A) while the character felt “suspicious” in (B).
b. The character felt “panicked” in (A) while the character felt “hopeful” in (B).
c. The character felt “impatient“ in (A) while the character felt “excited” in (B).
d. The character felt “satisfied” in (A) while the character felt “interested” in (B).

Q16. Harry and the older Dumbledore followed. She had a sharp-featured face that appeared more anxious than unkind, and she was talking over her shoulder to another aproned helper as she walked toward Dumbledore. ". and take the iodine upstairs to Martha, Billy Stubbs has been picking his scabs and Eric Whalley's oozing all over his sheets — chicken pox on top of everything else," she said to nobody in particular, and then her eyes fell upon Dumbledore and she stopped dead in her tracks, looking as astonished as if a giraffe had just crossed her threshold. "Good afternoon," said Dumbledore, holding out his hand. (A) "My name is Albus Dumbledore. I sent you a letter requesting an appointment and you very kindly invited me here today. " Apparently deciding that Dumbledore was not a hallucination, she said feebly, "Oh yes. Yes. " She led Dumbledore into a small room that seemed part sitting room, part office. She invited Dumbledore to sit on a rickety chair and seated herself behind a cluttered desk, eyeing him nervously. "I am here, as I told you in my letter, to discuss Tom Riddle and arrangements for his future," said Dumbledore. "Are you family?" "No, I am a teacher," said Dumbledore. "I have come to offer Tom a place at my school. " "What school's this, then?" "It is called Hogwarts," said Dumbledore. (B)"And how come you're interested in Tom?" "We believe he has qualities we are looking for. " "You mean he's won a scholarship? He's never been entered for one. " "Well, his name has been down for our school since birth — " "Who registered him? His parents?" Apparently Dumbledore thought so too, for Harry now saw him slip his wand out of the pocket of his velvet suit, at the same time picking up a piece of perfectly blank paper from

Miss Cole's desktop. "Here," said Dumbledore, waving his wand once as he passed her the piece of paper, "I think this will make everything clear. " "That seems perfectly in order," she said placidly, handing it back. "Er — may I offer you a glass of gin?" "Thank you very much," said Dumbledore, beaming. Smacking her lips frankly, she smiled at Dumbledore for the first time, and he didn't hesitate to press his advantage. "I was wondering whether you could tell me anything of Tom Riddle's history? I think he was born here in the orphanage?" "That's right," said Miss Cole, helping herself to more gin. "I remember it clear as anything, because I'd just started here myself. And she was dead in another hour. " "Did she say anything before she died?" asked Dumbledore. "Anything about the boy's father, for instance?" "Now, as it happens, she did," said Miss Cole, who seemed to be rather enjoying herself now, with the gin in her hand and an eager audience for her story.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “delighted” in (A) while the character felt “suspicious” in (B).
b. The character felt “uncomfortable” in (A) while the character felt “excited” in (B).
c. The character felt “frightened “ in (A) while the character felt “hopeful” in (B).
d. The character felt “enthusiastic” in (A) while the character felt “terrified” in (B).

Q17. "Time to go. " And Harry's feet left the floor to fall, seconds later, back onto the mg in front of Dumbledore's desk. "That's all there is?" said Harry blankly. Dumbledore had said that this was the most important memory of all, but he could not see what was so significant about it. (A) Admittedly the fog, and the fact that nobody seemed to have noticed it, was odd, but other than that nothing seemed to have happened except that Voldemort had asked a question and failed to get an answer. "As you might have noticed," said Dumbledore, reseating himself behind his desk, "that memory has been tampered with. " "Tampered with?" repeated Harry, sitting back down too. "Certainly," said Dumbledore. "Professor Slughorn has meddled with his own recollections. " "But why would he do that?" "Because, I think, he is ashamed of what he remembers," said Dumbledore. "He has tried to rework the memory to show himself in a better light, obliterating those parts which he does not wish me to see. "And so, for the first time, I am giving you homework, Harry. It will be your job to persuade Professor Slughorn to divulge the real memory, which will undoubtedly be our most crucial piece of information of all. " (B) Harry stared at him. "But surely, sir," he said, keeping his voice as respectful as possible, "you don't need me — you could use Legilimency or Veritaserum." "Professor Slughorn is an extremely able wizard who will be expecting both," said Dumbledore. "He is much more accomplished at Occlumency than poor Morfin Gaunt, and I would be astonished if he has not carried an antidote to Veritaserum with him ever since I coerced him into giving me this travesty of a recollection. "No, I think it would be foolish to attempt to wrest the truth from Professor Slughorn by force, and might do much more harm than good; I do not wish him to leave Hogwarts. It is most important that we secure the true memory, Harry. So, good luck . and good night. " A little taken

aback by the abrupt dismissal, Harry got to his feet quickly. "Good night, sir. " As he closed the study door behind him, he distinctly heard Phineas Nigellus say, "I can't see why the boy should be able to do it better than you, Dumbledore. "

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “confused” in (A) while the character felt “doubtful” in (B).
b. The character felt “amazed” in (A) while the character felt “confused” in (B).
c. The character felt “relieved” in (A) while the character felt “lucky” in (B).
d. The character felt “bored” in (A) while the character felt “powerless” in (B).

Q18. (A) Sure enough, the greenish light seemed to be growing larger at last, and within minutes, the boat had come to a halt, bumping gently into something that Harry could not see at first, but when he raised his illuminated wand he saw that they had reached a small island of smooth rock in the center of the lake. "Careful not to touch the water," said Dumbledore again as Harry climbed out of the boat. The island was no larger than Dumbledore 's office, an expanse of flat dark stone on which stood nothing but the source of that greenish light, which looked much brighter when viewed close to. Harry squinted at it; at first, he thought it was a lamp of some kind, but then he saw that the light was coming from a stone basin rather like the Pensieve, which was set on top of a pedestal. Dumbledore approached the basin and Harry followed. "What is it?" asked Harry quietly. "I am not sure," said Dumbledore. "Something more worrisome than blood and bodies, however. " Dumbledore pushed back the sleeve of his robe over his blackened hand, and stretched out the tips of his burned fingers toward the surface of the potion. "Sir, no, don't touch — !" "I cannot touch," said Dumbledore, smiling faintly. "See? You try. " Staring, Harry put his hand into the basin and attempted to touch the potion. "Out of the way, please, Harry," said Dumbledore. Harry remained silent while Dumbledore worked, but after a while Dumbledore withdrew his wand, and Harry felt it was safe to talk again. "You think the Horcrux is in there, sir?" "Oh yes. " Dumbledore peered more closely into the basin. Harry saw his face reflected, upside down, in the smooth surface of the green potion. "But how to reach it? This potion cannot be penetrated by hand, Vanished, parted, scooped up, or siphoned away, nor can it be Transfigured, Charmed, or otherwise made to change its nature. " Almost absentmindedly, Dumbledore raised his wand again, twirled it once in midair, and then caught the crystal goblet that he had conjured out of nowhere. "I can only conclude that this potion is supposed to be drunk. " "What?" said Harry. "No!" (B)"Yes, I think so: Only by drinking it can I empty the basin and see what lies in its depths. " "But what if — what if it kills you?" "Oh, I doubt that it would work like that," said Dumbledore easily. "Lord Voldemort would not want to kill the person who reached this island. " Harry couldn't believe it. Was this more of Dumbledore 's insane determination to see good in everyone? "Sir," said Harry, trying to keep his voice reasonable, "sir, this is Voldemort we're — " "I'm sorry, Harry; I should have said, he would not want to immediately kill the person who reached this island," Dumbledore corrected himself.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “amazed” in (A) while the character felt “determined” in (B).
b. The character felt “excited” in (A) while the character felt “scared” in (B).
c. The character felt “feared” in (A) while the character felt “angry” in (B).
d. The character felt “relieved” in (A) while the character felt “disappointed” in (B).

Q19. "Well, that's that plan scuppered," said George. "Obviously there's no chance at all of us getting a bit of your hair unless you cooperate. " "Yeah, thirteen of us against one bloke who's not allowed to use magic; we've got no chance," said Fred. "Funny," said Harry, "really amusing. " "If it has to come to force, then it will," growled Moody, his magical eye now quivering a little in its socket as he glared at Harry. "Everyone here's overage, Potter, and they're all prepared to take the risk. " Mundungus shrugged and grimaced; the magical eye swerved sideways to glare at him out of the side of Moody's head. "Let's have no more arguments. I want a few of your hairs, boy, now. " "But this is mad, there's no need — " "No need!" snarled Moody. "With You-Know-Who out there and half the Ministry on his side? (A) Potter, if we're lucky hell have swallowed the fake bait and he'll be planning to ambush you on the thirtieth, but he'd be mad not to have a Death Eater or two keeping an eye out, it's what I'd do. Even You-Know-Who can't split himself into seven. " Harry caught Hermione's eye and looked away at once. "So, Potter — some of your hair, if you please. " Harry glanced at Ron, who grimaced at him in a just- do-it sort of way. "Now!" barked Moody. With all of their eyes upon him, Harry reached up to the top of his head, grabbed a hank of hair, and pulled. "Good," said Moody, limping forward as he pulled the stopper out of the flask of potion. "Straight in here, if you please. " Harry dropped the hair into the mudlike liquid. (B) "Ooh, you look much tastier than Crabbe and Goyle, Harry," said Hermione, before catching sight of Ron's raised eyebrows, blushing slightly, and saying, "Oh, you know what I mean — Goyle 's potion looked like bogies. " "Right then, fake Potters line up over here, please," said Moody. Ron, Hermione, Fred, George, and Fleur lined up in front of Aunt Petunia's gleaming sink. "We're one short," said Lupin. "Here," said Hagrid gruffly, and he lifted Mundungus by the scruff of the neck and dropped him down beside Fleur, who wrinkled her nose pointedly and moved along to stand between Fred and George instead. "I've toldjer, I'd sooner be a protector," said Mundungus. "

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “rage” in (A) while the character felt “cheerful” in (B).
b. The character felt “destroyed” in (A) while the character felt “amused” in (B).
c. The character felt “disappointed” in (A) while the character felt “scared” in (B).
d. The character felt “painful” in (A) while the character felt “relieved” in (B).

Q20. "He was laughing?" said Harry in a hollow voice. "Oh yes," said Dumbledore. "You see, Kreacher was not able to betray us totally. He is not Secret-Keeper for the Order, he could not give the Malfoys our whereabouts or tell them any of the Order's confidential plans that he had been forbidden to reveal. He was bound by the enchantments of his kind, which is to say that he could not disobey a direct order from his master, Sirius. But he gave Narcissa information of the sort that is very valuable to Voldemort, yet must have seemed much too trivial for Sirius to think of banning him from repeating it. " "Like what?" said Harry. (A)"Like the fact that the person Sirius cared most about in the world was you," said Dumbledore quietly. "Like the fact that you were coming to regard Sirius as a mixture of father and brother. Voldemort knew already, of course, that Sirius was in the Order, that you knew where he was — but Kreacher's information made him realize that the one person whom you would go to any lengths to rescue was Sirius Black. " Harry's lips were cold and numb. "So . when I asked Kreacher if Sirius was there last night . " "The Malfoys — undoubtedly on Voldemort's instructions — had told him he must find a way of keeping Sirius out of the way once you had seen the vision of Sirius being tortured. Then, if you decided to check whether Sirius was at home or not, Kreacher would be able to pretend he was not. Kreacher injured Buckbeak the hippogriff yesterday, and at the moment when you made your appearance in the fire, Sirius was upstairs trying to tend to him. " There seemed to be very little air in Harry's lungs, his breathing was quick and shallow. "And Kreacher told you all this . and laughed?" "He did not wish to tell me," said Dumbledore. "But I am a sufficiently accomplished Legilimens myself to know when I am being lied to and I — persuaded him — to tell me the full story, before I left for the Department of Mysteries. " "And," whispered Harry, his hands curled in cold fists on his knees, "and Hermione kept telling us to be nice to him — " "She was quite right, Harry," said Dumbledore. (B) "I warned Sirius when we adopted twelve Grimmauld Place as our headquarters that Kreacher must be treated with kindness and respect. I also told him that Kreacher could be dangerous to us. I do not think that Sirius took me very seriously, or that he ever saw Kreacher as a being with feelings as acute as a humans — " "Don't you blame — don't you — talk — about Sirius like — " Harry's breath was constricted, he could not get the words out properly.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “convinced” in (A) while the character felt “regret” in (B).
b. The character felt “mad” in (A) while the character felt “disappointed” in (B).
c. The character felt “satisfied” in (A) while the character felt “forbidden” in (B).
d. The character felt “rage” in (A) while the character felt “painful” in (B).

Q21. "Hagrid was late for the start-of-term feast, just like Potter here, so I took it instead. (A) And incidentally," said Snape, standing back to allow Harry to pass him, "I was interested to see your new Patronus. " "I think you were better off with the old one," said Snape, the malice in his voice unmistakable. "The new one looks weak. " As Snape swung the lantern about, Harry saw,

fleetingly, a look of shock and anger on Tonks's face. "Good night," Harry called to her over his shoulder, as he began the walk up to the school with Snape. "Thanks for . everything. " "See you, Harry. " Snape did not speak for a minute or so. (B) Harry felt as though his body was generating waves of hatred so powerful that it seemed incredible that Snape could not feel them burning him. He had loathed Snape from their first encounter, but Snape had placed himself forever and irrevocably beyond the possibility of Harry's forgiveness by his attitude toward Sirius. Whatever Dumbledore said, Harry had had time to think over the summer, and had concluded that Snape 's snide remarks to Sirius about remaining safely hidden while the rest of the Order of the Phoenix were off fighting Voldemort had probably been a powerful factor in Sirius rushing off to the Ministry the night that he had died. Harry clung to this notion, because it enabled him to blame Snape, which felt satisfying, and also because he knew that if anyone was not sorry that Sirius was dead, it was the man now striding next to him in the darkness. "Fifty points from Gryffindor for lateness, I think," said Snape. "And, let me see, another twenty for your Muggle attire. You might have set a record, Potter. " The fury and hatred bubbling inside Harry seemed to blaze white-hot, but he would rather have been immobilized all the way back to London than tell Snape why he was late. "I suppose you wanted to make an entrance, did you?" Snape continued. "And with no flying car available you decided that bursting into the Great Hall halfway through the feast ought to create a dramatic effect. " Still Harry remained silent, though he thought his chest might explode. He knew that Snape had come to fetch him for this, for the few minutes when he could needle and torment Harry without anyone else listening. Harry wondered whether he could slip his Invisibility Cloak back on, thereby gaining his seat at the long Gryffindor table (which, inconveniently, was the farthest from the entrance hall) without being noticed. As though he had read Harry's mind, however, Snape said, "No cloak. You can walk in so that everyone sees you, which is what you wanted, I'm sure. " Harry turned on the spot and marched straight through the open doors: anything to get away from Snape.

How is the main characters sentiment described in (A) and (B) different?

a. The character felt “enjoying” in (A) while the character felt “furious” in (B).
b. The character felt “suspicious” in (A) while the character felt “rage” in (B).
c. The character felt “fun” in (A) while the character felt “weak” in (B).
d. The character felt “enthusiastic” in (A) while the character felt “feared” in (B).

A4: Items from Divergent Item Model 2

Q22. Mister Dursley hummed as he picked out his most boring tie for work, and Miss Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair. At half past eight, Mister Dursley picked up his briefcase, pecked Miss Dursley on the cheek, and tried to kiss Dudley good-bye but missed, because Dudley was now having a tantrum and throwing his cereal at the walls. "Little tyke," chortled Mister Dursley as he left the house. For a second, Mister Dursley didn't realize what he had seen — then he jerked his head around to look again. Mister Dursley blinked and stared at the cat. As Mister Dursley drove around the corner and up the road, he watched the cat in his mirror. Mister Dursley gave himself a little shake and put the cat out of his mind. Mister Dursley couldn't bear people who dressed in funny clothes — the getups you saw on young people! Mister Dursley was enraged to see that a couple of them weren't young at all; why, that man had to be older than he was, and wearing an emerald-green cloak! But then it struck Mister Dursley that this was probably some silly stunt — these people were obviously collecting for something . yes, that would be it. The traffic moved on and a few minutes later, Mister Dursley arrived in the Grunnings parking lot, his mind back on drills. Mister Dursley always sat with his back to the window in his office on the ninth floor. Mister Dursley, however, had a perfectly normal, owl-free morning. " The Potters, that's right, that's what I heard — " " — yes, their son, Harry — " Mister Dursley stopped dead. Potter wasn't such an unusual name. He was sure there were lots of people called Potter who had a son called Harry. Come to think of it, he wasn't even sure his nephew was called Harry. "Sorry," he grunted, as the tiny old man stumbled and almost fell. It was a few seconds before Mister Dursley realized that the man was wearing a violet cloak. On the contrary, his face split into a wide smile and he said in a squeaky voice that made passersby stare, "Don't be sorry, my dear sir, for nothing could upset me today! Even Muggles like yourself should be celebrating, this happy, happy day!" And the old man hugged Mister Dursley around the middle and walked off. Mister Dursley stood rooted to the spot. "Shoo!" said Mister Dursley loudly. Mister Dursley wondered. She told him over dinner all about Miss Next Door's problems with her daughter and how Dudley had learned a new word ("Won't!"). Mister Dursley tried to act normally.

Which of the following indicates different sentiment of the characters from the others?

a. “The Potters, that's right, that's what I heard — yes, their son, Harry — “
b. “Mister Dursley, however, had a perfectly normal, owl-free morning.”
c. “Little tyke," chortled Mister Dursley as he left the house.”
d. “Mister Dursley hummed as he picked out his most boring tie for work, and Miss Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair.”

Q23. Filch grabbed a quill from a pot on his desk and began shuffling around looking for parchment. "Dung," he muttered furiously, "great sizzling dragon bogies . frog brains . rat intestines . I've had enough of it . make an example . where's the form . yes . " He retrieved a

large roll of parchment from his desk drawer and stretched it out in front of him, dipping his long black quill into the ink pot. " Name . Harry Potter. Crime . " "It was only a bit of mud!" said Harry. "It's only a bit of mud to you, boy, but to me it's an extra hour scrubbing!" shouted Filch, a drip shivering unpleasantly at the end of his bulbous nose. "Crime . befouling the castle . suggested sentence . " Dabbing at his streaming nose, Filch squinted unpleasantly at Harry, who waited with bated breath for his sentence to fall. But as Filch lowered his quill, there was a great BANG! "PEEVES!" Filch roared, flinging down his quill in a transport of rage. "I'll have you this time, I'll have you!" And without a backward glance at Harry, Filch ran flat-footed from the office, Miss Norris streaking alongside him. Harry didn't much like Peeves, but couldn't help feeling grateful for his timing. Hopefully, whatever Peeves had done (and it sounded as though he'd wrecked something very big this time) would distract Filch from Harry. Thinking that he should probably wait for Filch to come back, Harry sank into a moth-eaten chair next to the desk. With a quick glance at the door to check that Filch wasn't on his way back, Harry picked up the envelope and read: KWIKSPELL A Correspondence Course in Beginners' Magic Intrigued, Harry flicked the envelope open and pulled out the sheaf of parchment inside. Madam Z. Nettles of Topsham writes: "I had no memory for incantations and my potions were a family joke! Now, after a Kwikspell course, I am the center of attention at parties and friends beg for the recipe of my Scintillation Solution!" Prod of Didsbury says: "My wife used to sneer at my feeble charms, but one month into your fabulous Kwikspell course and I succeeded in turning her into a yak! Thank you, Kwikspell!" Fascinated, Harry thumbed through the rest of the envelope's contents. Why on earth did Filch want a Kwikspell course? Harry was just reading "Lesson One: Holding Your Wand (Some Useful Tips)" when shuffling footsteps outside told him Filch was coming back. Stuffing the parchment back into the envelope, Harry threw it back onto the desk just as the door opened. Filch was looking triumphant. "That vanishing cabinet was extremely valuable!"

Which of the following indicates different sentiment of the characters from the others?

a. "Harry didn't much like Peeves, but couldn't help feeling grateful for his timing."
b. “Filch roared, flinging down his quill in a transport of rage.”
c. "It's only a bit of mud to you, boy, but to me it's an extra hour scrubbing!" shouted Filch, a drip shivering unpleasantly at the end of his bulbous nose.
d. Dabbing at his streaming nose, Filch squinted unpleasantly at Harry, who waited with bated breath for his sentence to fall.

Q24. Just for background reading. " "But the thing is, it's in the Restricted Section of the library, so I need a teacher to sign for it — I'm sure it would help me understand what you say in Gadding with Ghouls about slow-acting venoms — " "Ah, Gadding with GhoulsV' said Lockhart, taking the note from Hermione and smiling widely at her. "Possibly my very favorite book. You enjoyed it?" "Oh, yes," said Hermione eagerly. "So clever, the way you trapped that

last one with the tea-strainer — " "Well, I'm sure no one will mind me giving the best student of the year a little extra help," said Lockhart warmly, and he pulled out an enormous peacock quill. "Yes, nice, isn't it?" he said, misreading the revolted look on Ron's face. "I usually save it for book signings. " He scrawled an enormous loopy signature on the note and handed it back to Hermione. "So, Harry," said Lockhart, while Hermione folded the note with fumbling fingers and slipped it into her bag. "Tomorrow's the first Quidditch match of the season, I believe? Gryffindor against Slytherin, is it not? . " Harry made an indistinct noise in his throat and then hurried off after Ron and Hermione. "I don't believe it," he said as the three of them examined the signature on the note. "He didn't even look at the book we wanted. " "That's because he's a brainless git," said Ron. "But who cares, we've got what we needed — " "He is not a brainless git," said Hermione shrilly as they half ran toward the library. "Just because he said you were the best student of the year — " They dropped their voices as they entered the muffled stillness of the library. "Moste Potente Potions?" she repeated suspiciously, trying to take the note from Hermione; but Hermione wouldn't let go. "I was wondering if I could keep it," she said breathlessly. "Oh, come on," said Ron, wrenching it from her grasp and thrusting it at Madam Pince. "We'll get you another autograph. Lockhart 'll sign anything if it stands still long enough. " Hermione put it carefully into her bag and they left, trying not to walk too quickly or look too guilty. Five minutes later, they were barricaded in Moaning Myrtle's out-of-order bathroom once again. Hermione had overridden Ron's objections by pointing out that it was the last place anyone in their right minds would go, so they were guaranteed some privacy. Moaning Myrtle was crying noisily in her stall, but they were ignoring her, and she them. Hermione opened Moste Potente Potions carefully, and the three of them bent over the damp-spotted pages. "Here it is," said Hermione excitedly as she found the page headed The Polyjuice Potion. Harry sincerely hoped the artist had imagined the looks of intense pain on their faces.

Which of the following indicates different sentiment from the others?

a. “So clever, the way you trapped that last one with the tea-strainer —”
b. “Well, I'm sure no one will mind me giving the best student of the year a little extra help,”
c. “Yes, nice, isn't it?" he said, misreading the revolted look on Ron's face. “I usually save it for book signings.”
d. “Hermione put it carefully into her bag and they left, trying not to walk too quickly or look too guilty.”

Q25. Could he be a descendant of Salazar Slytherin? The Dursleys had always forbidden questions about his wizarding relatives. Quietly, Harry tried to say something in Parseltongue. But I'm in Gryffindor, Harry thought. The Sorting Hat wouldn't have put me in here if I had Slytherin blood. Ah, said a nasty little voice in his brain, but the Sorting Hat wanted to put you in Slytherin, don't you remember? Harry turned over. He'd see Justin the next day in Herbology and he'd explain that he'd been calling the snake off, not egging it on, which (he thought angrily,

pummeling his pillow) any fool should have realized. By next morning, however, the snow that had begun in the night had turned into a blizzard so thick that the last Herbology lesson of the term was canceled: Professor Sprout wanted to fit socks and scarves on the Mandrakes, a tricky operation she would entrust to no one else, now that it was so important for the Mandrakes to grow quickly and revive Miss Norris and Colin Creevey. Harry fretted about this next to the fire in the Gryffindor common room, while Ron and Hermione used their time off to play a game of wizard chess. "For heaven's sake, Harry," said Hermione, exasperated, as one of Ron's bishops wrestled her knight off his horse and dragged him off the board. "Go and find Justin if it's so important to you. " So Harry got up and left through the portrait hole, wondering where Justin might be. Shivering, Harry walked past classrooms where lessons were taking place, catching snatches of what was happening within. Professor McGonagall was shouting at someone who, by the sound of it, had turned his friend into a badger. Resisting the urge to take a look, Harry walked on by, thinking that Justin might be using his free time to catch up on some work, and deciding to check the library first. Between the long lines of high bookshelves, Harry could see that their heads were close together and they were having what looked like an absorbing conversation. He couldn't see whether Justin was among them. "So anyway," a stout boy was saying, "I told Justin to hide up in our dormitory. I mean to say, if Potter's marked him down as his next victim, it's best if he keeps a low profile for a while. Of course, Justin's been waiting for something like this to happen ever since he let slip to Potter he was Muggle-born. Justin actually told him he'd been down for Eton. That's not the kind of thing you bandy about with Slytherin's heir on the loose, is it?" "You definitely think it is Potter, then, Ernie?" "Hannah," said the stout boy solemnly, "he's a Parselmouth. They called Slytherin himself Serpent-tongue. " There was some heavy murmuring at this, and Ernie went on, "Remember what was written on the wall?

Which of the following indicates different sentiment from the others?

a. “Dursleys had always forbidden questions about his wizarding relatives.”
b. “Ah, said a nasty little voice in his brain, but the Sorting Hat wanted to put you in Slytherin, don't you remember?”
c. “Harry got up and left through the portrait hole, wondering where Justin might be. Shivering,”
d. **“Hannah, said the stout boy solemnly, he's a Parselmouth. They called Slytherin himself Serpent-tongue.”**

Q26. Hermione looked aghast at the news. "But — only a Gryffindor could have stolen — nobody else knows our password — " "Exactly," said Harry. "Perfect Quidditch conditions!" said Wood enthusiastically at the Gryffindor table, loading the team's plates with scrambled eggs. "Harry, buck up there, you need a decent breakfast. " Harry had been staring down the packed Gryffindor table, wondering if the new owner of Riddle's diary was right in front of his eyes. Hermione had been urging him to report the robbery, but Harry didn't like the idea. He'd

have to tell a teacher all about the diary, and how many people knew why Hagrid had been expelled fifty years ago? As he left the Great Hall with Ron and Hermione to go and collect his Quidditch things, another very serious worry was added to Harry's growing list. He had just set foot on the marble staircase when he heard it yet again — " Kill this time . let me rip . tear . " He shouted aloud and Ron and Hermione both jumped away from him in alarm. "The voice!" said Harry, looking over his shoulder. "I just heard it again — didn't you?" Ron shook his head, wide-eyed. Hermione, however, clapped a hand to her forehead. "Harry — I think I've just understood something! I've got to go to the library!" And she sprinted away, up the stairs. " What does she understand?" said Harry distractedly, still looking around, trying to tell where the voice had come from. "Loads more than I do," said Ron, shaking his head. "But why's she got to go to the library?" "Because that's what Hermione does," said Ron, shrugging. "When in doubt, go to the library. " Harry stood, irresolute, trying to catch the voice again, but people were now emerging from the Great Hall behind him, talking loudly, exiting through the front doors on their way to the Quidditch pitch. "You'd better get moving," said Ron. "It's nearly eleven — the match — " Harry raced up to Gryffindor Tower, collected his Nimbus Two Thousand, and joined the large crowd swarming across the grounds, but his mind was still in the castle along with the bodiless voice, and as he pulled on his scarlet robes in the locker room, his only comfort was that everyone was now outside to watch the game. Harry was just mounting his broom when Professor McGonagall came half marching, half running across the pitch, carrying an enormous purple megaphone. Harry's heart dropped like a stone. "This match has been canceled," Professor McGonagall called through the megaphone, addressing the packed stadium. Oliver Wood, looking devastated, landed and ran toward Professor McGonagall without getting off his broomstick. "But, Professor!"

Which of the following indicates different sentiment from the others?

a. “But — only a Gryffindor could have stolen — nobody else knows our password —”
b. “Harry had been staring down the packed Gryffindor table, wondering if the new owner of Riddle's diary was right in front of his eyes.”
c. “He'd have to tell a teacher all about the diary, and how many people knew why Hagrid had been expelled fifty years ago?”
d. "Perfect Quidditch conditions!, said Wood enthusiastically”

Q26. "What's that for?" said Harry, pointing at the crossbow as they stepped inside. "Nothin' — nothin' — " Hagrid muttered. "I've bin expectin' — doesn' matter — Sit down — I'll make tea He hardly seemed to know what he was doing. "Are you okay, Hagrid?" said Harry. "Did you hear about Hermione?" "Oh, I heard, all righ'," said Hagrid, a slight break in his voice. Hagrid dropped the fruitcake. Harry and Ron exchanged panic-stricken looks, then threw the Invisibility Cloak back over themselves and retreated into a corner. Hagrid checked that they were hidden, seized his crossbow, and flung open his door once more. "Good evening, Hagrid. " It was

Dumbledore. "That's Dad's boss!" Ron breathed. "Cornelius Fudge, the Minister of Magic!" Harry elbowed Ron hard to make him shut up. Hagrid had gone pale and sweaty. He dropped into one of his chairs and looked from Dumbledore to Cornelius Fudge. "Bad business, Hagrid," said Fudge in rather clipped tones I never," said Hagrid, looking imploringly at Dumbledore. "You know I never, Professor Dumbledore, sir — " "I want it understood, Cornelius, that Hagrid has my full confidence," said Dumbledore, frowning at Fudge. "Look, Albus," said Fudge, uncomfortably. "Hagrid's record's against him. Ministry's got to do something — the school governors have been in touch — " "Yet again, Cornelius, I tell you that taking Hagrid away will not help in the slightest," said Dumbledore. His blue eyes were full of a fire Harry had never seen before. "Look at it from my point of view," said Fudge, fidgeting with his bowler. "I'm under a lot of pressure. If it turns out it wasn't Hagrid, he'll be back and no more said. Wouldn't be doing my duty — " "Take me?" said Hagrid, who was trembling. "Take me where?" "For a short stretch only," said Fudge, not meeting Hagrid 's eyes. "Not a punishment, Hagrid, more a precaution. If someone else is caught, you'll be let out with a full apology — " "Not Azkaban?" croaked Hagrid. Dumbledore answered it. It was Harry's turn for an elbow in the ribs; he'd let out an audible gasp. Mister Lucius Malfoy strode into Hagrid's hut, swathed in a long black traveling cloak, smiling a cold and satisfied smile. Fang started to growl. "Already here, Fudge," he said approvingly. "Good, good. " "What're you doin' here?" said Hagrid furiously. "Get outta my house!" "My dear man, please believe me, I have no pleasure at all in being inside your — er — d'you call this a house?" said Lucius Malfoy, sneering as he looked around the small cabin. "I simply called at the school and was told that the headmaster was here. " "And what exactly did you want with me, Lucius?" said Dumbledore. He spoke politely, but the fire was still blazing in his blue eyes. "

Which of the following indicates different sentiment from the others?

a. “Not a punishment, Hagrid, more a precaution. If someone else is caught, you'll be let out with a full apology —”
b. “Mister Lucius Malfoy strode into Hagrid's hut, swathed in a long black traveling cloak, smiling a cold and satisfied smile.”
c. “My dear man, please believe me, I have no pleasure at all in being inside your — er — d'you call this a house?”
d. “Already here, Fudge," he said approvingly. "Good, good. ”*

Q27. Fang flung himself at the window in his anxiety to get out, and when Harry opened the door, he shot off through the trees to Hagrid's house, tail between his legs. Harry got out too, and after a minute or so, Ron seemed to regain the feeling in his limbs and followed, still stiff-necked and staring. Harry gave the car a grateful pat as it reversed back into the forest and disappeared from view. Harry went back into Hagrid's cabin to get the Invisibility Cloak. Fang was trembling under a blanket in his basket. When Harry got outside again, he found Ron being violently sick

in the pumpkin patch. "Follow the spiders," said Ron weakly, wiping his mouth on his sleeve. "I'll never forgive Hagrid. We're lucky to be alive. " "I bet he thought Aragog wouldn't hurt friends of his," said Harry. "That's exactly Hagrid's problem!" said Ron, thumping the wall of the cabin. "He always thinks monsters aren't as bad as they're made out, and look where it's got him! A cell in Azkaban!" "What was the point of sending us in there? What have we found out, I'd like to know?" "That Hagrid never opened the Chamber of Secrets," said Harry, throwing the cloak over Ron and prodding him in the arm to make him walk. "He was innocent. " Ron gave a loud snort. Evidently, hatching Aragog in a cupboard wasn't his idea of being innocent. As the castle loomed nearer Harry twitched the cloak to make sure their feet were hidden, then pushed the creaking front doors ajar. At last they reached the safety of the Gryffindor common room, where the fire had burned itself into glowing ash. Ron fell onto his bed without bothering to get undressed. Harry, however, didn't feel very sleepy. He sat on the edge of his fourposter, thinking hard about everything Aragog had said. The creature that was lurking somewhere in the castle, he thought, sounded like a sort of monster Voldemort — even other monsters didn't want to name it. But he and Ron were no closer to finding out what it was, or how it Petrified its victims. Even Hagrid had never known what was in the Chamber of Secrets. Harry swung his legs up onto his bed and leaned back against his pillows, watching the moon glinting at him through the tower window. Riddle had caught the wrong person, the Heir of Slytherin had got off, and no one could tell whether it was the same person, or a different one, who had opened the Chamber this time. Harry lay down, still thinking about what Aragog had said. "Ron," he hissed through the dark, "Ron — " Ron woke with a yelp like Fang's, stared wildly around, and saw Harry. "Ron — that girl who died. Aragog said she was found in a bathroom," said Harry, ignoring Neville's snuffling snores from the corner. "What if she never left the bathroom? What if she's still there?"

Which of the following indicates different sentiment from the others?

a. “When Harry got outside again, he found Ron being violently sick in the pumpkin patch.”*
b. “I'll never forgive Hagrid. We're lucky to be alive.”
c. “That's exactly Hagrid's problem!" said Ron, thumping the wall of the cabin.”
d. "He always thinks monsters aren't as bad as they're made out, and look where it's got him! A cell in Azkaban!"

Q28. "All right, I've got the point," said Harry. "Well, we won't find out unless we look at it," he said, and he ducked around Ron and picked it up off the floor. Harry saw at once that it was a diary, and the faded year on the cover told him it was fifty years old. On the first page he could just make out the name "T. M. Riddle" in smudged ink. "Hang on," said Ron, who had approached cautiously and was looking over Harry's shoulder. "I know that name. . T. M. Riddle got an award for special services to the school fifty years ago. " "How on earth d'you know that?" said Harry in amazement. "Because Filch made me polish his shield about fifty times in

detention," said Ron resentfully. "That was the one I burped slugs all over. If you'd wiped slime off a name for an hour, you'd remember it, too. " Harry peeled the wet pages apart. "He never wrote in it," said Harry, disappointed. "I wonder why someone wanted to flush it away?" said Ron curiously. Harry turned to the back cover of the book and saw the printed name of a variety store on Vauxhall Road, London. "He must've been Muggle-born," said Harry thoughtfully. "To have bought a diary from Vauxhall Road. . " "Well, it's not much use to you," said Ron. "Fifty points if you can get it through Myrtle's nose. " Harry, however, pocketed it. Hermione left the hospital wing, de-whiskered, tail- less, and fur-free, at the beginning of February. On her first evening back in Gryffindor Tower, Harry showed her T. M. Riddle's diary and told her the story of how they had found it. "Oooh, it might have hidden powers," said Hermione enthusiastically, taking the diary and looking at it closely. "If it has, it's hiding them very well," said Ron. "Maybe it's shy. I don't know why you don't chuck it, Harry. " "I wish I knew why someone did try to chuck it," said Harry. "I wouldn't mind knowing how Riddle got an award for special services to Hogwarts either. " "Could've been anything," said Ron. "Maybe he got thirty O. W. L. s or saved a teacher from the giant squid. Maybe he murdered Myrtle; that would've done everyone a favor. . " But Harry could tell from the arrested look on Hermione 's face that she was thinking what he was thinking. "What?" said Ron, looking from one to the other. "Well, the Chamber of Secrets was opened fifty years ago, wasn't it?" "That's what Malfoy said. " "Yeah . " said Ron slowly. "And this diary is fifty years old," said Hermione, tapping it excitedly. "So?" "Oh, Ron, wake up," snapped Hermione. "We know the person who opened the Chamber last time was expelled fifty years ago. We know T. M. Riddle got an award for special services to the school fifty years ago.

Which of the following indicates different sentiment from the others?

a. “I wouldn't mind knowing how Riddle got an award for special services to Hogwarts either.”
b. “Maybe he murdered Myrtle; that would've done everyone a favor.”
c. “Oooh, it might have hidden powers,” said Hermione”
d. “He never wrote in it,” said Harry”

Appendix B

B1: Example Python Code of Item Model 1

```
def item_model1(data = coherent, df_wrd=df_wrd, num= i): # takes the sample corpus, topic-word distribution

    # check whether the text was inappropriately parsed in the middle of the sentence
    text = data.samples.iloc[num]
    try:
        text = clean_text(text)
    except:
        text = text

    q=[] # saves all the test item variations

    topic_index = data.drop(columns=['chapter', 'total', 'samples', 'dominant', 'length']).iloc[num]
    topicA = topic_index.argmax()+1
    words = df_wrd.iloc[topicA-1]
    keyword_list = words[words!=0].index

    for i in [k for k in keyword_list if k in nltk.word_tokenize(text.lower())]:
        topicAkeyword = i

       # specifies the stem with the topic keyword
        stem= "The main character's feeling " + "" +str(topicAkeyword) + "" +  " is most likely related to the statement:
"

      # specifies the topic sentence which has highest corresponding topic weight in the given text.
        keyed_option = sent_score(text, topicA)

       # distractors are located by identifying topic sentences from the rest of the topics.
        distractors = []
        for i in range(1, 11):
            distractors.append(sent_score(text, i))
        distractors= [sub for lst in distractors for sub in lst]
        distractors = [i for i in list(set(distractors)) if i not in keyed_option]
        q.append((text, stem, keyed_option, distractors))

    return q
```

B2: Example Python Code of Item Model 4

```
def item_model4(data = divergent, df_wrd=df_wrd, num= i):

    text = data.samples.iloc[num]
    text = clean_text(text)
    topic_index =data.drop(columns=['chapter', 'total', 'samples', 'dominant','length']).iloc[num]

    q = []
    for k in topic_index[topic_index!=0].index:
        topicA = int(k.split('. ')[-1])


        stem= "Which of the following indicates different sentiment from the others?"
        print('Stem: ', stem)

        keyed_optionA = sent_score(text, topicA)
        distractors = [sent_score(text, i) for i in range(1,11) if i != topicA]
        distractors= list(set([sub for lst in distractors for sub in lst]))
        distractors= [i for i in distractors if i not in keyed_optionA]

        q1 = [text, stem, keyed_optionA, distractors]
        print('---- Topic A is Topic: ', topicA)
        print('---- Topic B is Topic: ', topicB)

        stem= "Which of the following indicates different sentiment from the others?"
        print('Stem: ', stem)

        keyed_optionA = sent_score(text, topicA)
        distractors = [sent_score(text, i) for i in range(1,11) if i != topicA]
        distractors= list(set([sub for lst in distractors for sub in lst]))
        distractors= [i for i in distractors if i not in keyed_optionA]
        q.append((text, stem, keyed_optionA, distractors))
    return q
```