

**Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing
Education**

By

Tracey C. Stephen

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

Doctor of Philosophy

in

MEASUREMENT, EVALUATION, AND COGNITION

Department of Educational Psychology

University of Alberta

© Tracey C. Stephen, 2019

Abstract

Automated essay scoring (AES) is a developing technology that is increasingly recognized as a potential strategy for managing the challenges associated with testing and scoring written assessments. The importance of using open-ended writing assessments to facilitate higher-level thinking, including making connections and critical thinking, has been shown in several studies. Health sciences education fields are increasingly recognizing the importance of using essay-type response examinations to assess the performance of learners particularly in the areas of critical thinking, clinical reasoning, and clinical judgement. The areas of critical thinking, clinical reasoning, and clinical judgement are recognized as critical aspects of clinical practice affecting patient safety, but are difficult to accurately assess using only selected-response item format (such as multiple-choice questions) examinations. The complexity of assessing the areas of critical thinking, clinical reasoning, and clinical judgement in patient situations supports the inclusion of constructed-response items (such as short answer essay questions) in assessments. However, there are several challenges to using essay-type response examinations, including time and costs of scoring, consistency in scoring, marker fatigue, timely feedback, and impact of subjectivity. The following research uses AES to score a constructed-response item to assess critical thinking, clinical reasoning, and clinical judgement for nursing students. The focus of this study is limited to scoring written assessments and the primary purpose of this study is to evaluate the effectiveness of using AES to score constructed-response items to assess higher-level thinking skills in nursing education.

Keywords: Health sciences education, computerized assessment, feedback, patient safety, nursing education, constructed-response item, critical thinking, clinical judgement, clinical reasoning

Preface

This thesis is an original work by Tracey C. Stephen. The research project included in this thesis received research ethics approval from the University of Alberta Research Ethics Board, “Using Automated Essay Scoring to Assess Critical Thinking in Nursing Education”, No. Pro00071611, March 2, 2017.

An amendment to the original proposal was approved by the University of Alberta Research Ethics Board on June 29, 2018, No. Pro00071611_AME1.

Dedication

To Andy, Jenelle, Taylir, and Mackenna.

Thank you for everything you do and everything you are.

I love you more than I can ever tell you.

Acknowledgments

Thank you to Dr. Sharla King and Dr. Mark Gierl for your support, guidance, and brilliance. It is truly a privilege to work with you.

Thank you also to my committee members, Dr. Oksana Babenko, Dr. Okan Bulut, Dr. Mary Roberts, and Dr. Saad Chahine for sharing your expertise, support, and brilliance too.

Thank you to Dr. Craig Jamieson and Human Rater #2 for your expertise and willingness to try an adventure with me.

Thank you to my colleagues and friends for all your encouragement and support.

Finally, thank you to the students (past and present) who I am fortunate to work with. You are inspiring and encourage me to keep reaching.

Thank you

Table of Contents

Abstract.....	ii
Acknowledgments	iii
List of Tables	xi
List of Figures	xiii
Chapter 1: Introduction	1
Background of the Issue and Target Audience.....	4
Focus of the Research Study and Research Questions.....	7
Importance of this Research Study.....	8
Dissertation.....	9
Chapter 2: Literature Review.....	10
Section One: Assessing Higher-Level Thinking Skills in Nursing Education.....	10
Critical Thinking.....	11
Critical Thinking in Nursing Practice and Dual Process Theory.....	12
Critical Thinking Theoretical Framework: A Two-Phase Framework.....	14
Critical Thinking: A Two-Phase Framework Applied to a Patient Scenario.....	15
Critical Thinking: A Two-Phase Framework Applied to this Research Study.....	17
Clinical Reasoning and Clinical Judgement in Nursing Practice.....	18
Critical Thinking, Clinical Reasoning, and Clinical Judgement in Nursing Education...	19
Context-based learning with patient scenarios.....	19
Concept-mapping.....	20
Simulation.....	20
Virtual Simulation.....	21
Reflective Journaling.....	21
Critical Thinking and Patient Scenarios to Assess Learning.....	22
Feedback and Learning.....	23
Selected-Response and Constructed-Response Item Formats.....	25
Selected-Response Item Formats.....	25
Constructed-Response Item Formats.....	26

AUTOMATED ESSAY SCORING	vii
Section Two: Automated Essay Scoring.....	27
Automated Essay Scoring System Processes.....	27
Data Preparation.....	28
Feature Extraction.....	28
Machine Learning and Model Building.....	29
Scoring.....	32
WEKA—An Automated Essay Scoring System.....	32
LightSIDE.....	35
Advantages of Automated Essay Scoring.....	36
Human Raters and Automated Essay Scoring.....	37
Disadvantages of Automated Essay Scoring.....	37
Challenges to Automated Essay Scoring.....	38
Literature Review Conclusions.....	40
Chapter 3: Methods and Data Collection.....	41
Section One: Methods of Data Selection and Collection.....	41
Constructed- Response Items.....	41
Item Development.....	42
Ethics Approval.....	44
Selection of Data.....	45
Data Collection Procedures.....	46
Section Two: Analysis Procedures.....	49
Reliability Coefficients.....	49
Agreement Measures.....	50
Conclusion.....	51
Chapter 4: Results.....	53
Section One: Overview of Data Collection Processes and Export of Data Files.....	53
Scoring the Responses by the Second Human Rater.....	57
Section Two: Results from Automated Essay Scoring Program.....	61
Conversion and Upload Processes.....	61

Data Preprocessing and Python.....	65
Data preprocessing.....	66
Development of the AES Model.....	67
Section Three: Results from Comparisons between Human Raters and AES.....	68
Agreement Measures and Reliability Coefficients.....	68
Analysis of Comparisons for Item #1.....	70
Agreement measures and reliability coefficients.....	70
Score discrepancy analysis.....	72
Analysis of Comparisons for Item #2.....	74
Agreement measures and reliability coefficients.....	74
Score discrepancy analysis.....	76
Analysis of Comparisons for Item #3.....	78
Agreement measures and reliability coefficients.....	78
Score discrepancy analysis.....	81
Analysis of Comparisons for Item #4.....	82
Agreement measures and reliability coefficients.....	82
Score discrepancy analysis.....	84
Applying the AES Model.....	86
Section Four: Results from Analysis of Application of AES Model.....	87
Categories for Performance.....	87
Analysis of Comparisons for Item #1—Application of AES Model.....	89
Agreement measure and reliability coefficients.....	89
Score discrepancy analysis.....	91
Analysis of Comparisons for Item #2—Application of AES Model.....	93
Agreement measure and reliability coefficients.....	93
Score discrepancy analysis.....	95
Analysis of Comparisons for Item #3—Application of AES Model.....	97
Agreement measure and reliability coefficients.....	97
Score discrepancy analysis.....	97
Analysis of Comparisons for Item #4—Application of AES Model.....	99
Agreement measure and reliability coefficients.....	100

Score discrepancy analysis.....	101
Section Five: Summary and Discussion of Results.....	102
Comparisons Between Human Raters.....	103
Comparisons Between HR1 and AES.....	104
Comparisons Between HR2 and AES.....	106
Summary of Analysis and Discussion for Item #1.....	107
Summary of Analysis and Discussion for Item #2.....	109
Summary of Analysis and Discussion for Item #3.....	110
Summary of Analysis and Discussion for Item #4.....	111
Conclusion.....	112
Chapter 5: Discussion and Conclusion.....	114
Section One: Restatement of Research Questions and Summary of Methods and Results.....	116
Summary of Results.....	119
Research Question 1.....	119
Research Question 2.....	122
Research Question 3.....	122
Section Two: Limitations of this Research Study.....	125
Section Three: Directions for Future Research.....	129
Conclusions.....	131
References.....	133
Appendix A: Nursing and Health Sciences programs approached to access constructed-response items.....	150
Appendix B: Constructed-response Item for Automated Essay Scoring Analysis.....	151
Appendix C: Scoring Rubric Development—Responses from Colleagues.....	152
Appendix D: Scoring Rubric for Constructed-response Item for Automated Essay Scoring Analysis.....	161
Appendix E: Ethics Approval Application and Approval.....	163

Appendix F: Permission to Access Student Responses from the Faculty of Nursing, U of A.....	164
Appendix G: Code for Data Preprocessing for Python Developed and Used with Permission by Shin (2018).....	167
Appendix H: Ethics Application and Approval for Amendment to Study.....	175
Appendix I: Permission Letter to Vice Dean, Faculty of Nursing to Apply AES Model.....	176

List of Tables

Table 4.1. Agreement Measures and Reliability Coefficients for Item #1.....	71
Table 4.2. Score Discrepancy Analysis for Item #1.....	73
Table 4.3. Agreement Measures and Reliability Coefficients for Item #2.....	75
Table 4.4. Score Discrepancy Analysis for Item #2.....	77
Table 4.5. Agreement Measures and Reliability Coefficients for Item #3.....	79
Table 4.6. Score Discrepancy Analysis for Item #3.....	81
Table 4.7. Agreement Measures and Reliability Coefficients for Item #4.....	83
Table 4.8. Score Discrepancy Analysis for Item #4.....	85
Table 4.9. Agreement Measures and Reliability Coefficients for Item #1—Application of AES Model.....	90
Table 4.10. Score Discrepancy Analysis for Item #1—Application of AES Model.....	92
Table 4.11. Agreement Measures and Reliability Coefficients for Item #2—Application of AES Model.....	94
Table 4.12. Score Discrepancy Analysis for Item #2—Application of AES Model.....	96
Table 4.13. Agreement Measures and Reliability Coefficients for Item #3—Application of AES Model.....	97
Table 4.14. Score Discrepancy Analysis for Item #3—Application of AES Model.....	98
Table 4.15. Agreement Measures and Reliability Coefficients for Item #4—Application of AES Model.....	100
Table 4.16. Score Discrepancy Analysis for Item #4—Application of AES Model.....	101
Table 4.17. Summary of Comparisons between Human Raters.....	103

List of Tables (cont'd)

Table 4.18. Summary of Comparisons between HR1 and AES.....	106
Table 4.19. Summary of Comparisons between HR2 and AES.....	106
Table 4.20. Summary of Results for Item #1.....	109
Table 4.21. Summary of Results for Item #2.....	110
Table 4.22. Summary of Results for Item #3.....	111
Table 4.23. Summary of Results for Item #4.....	112

List of Figures

Figure 4.1. Example of responses for item #1 scored by HR1.....	54
Figure 4.2. Example of responses for item #2 scored by HR1.....	55
Figure 4.3. Example of responses for item #3 scored by HR1.....	56
Figure 4.4. Example of responses for item #4 scored by HR1.....	57
Figure 4.5. Example of scores for item #1 by case number for HR1 and HR2.....	59
Figure 4.6. Example of scores for item # 2 by case number for HR1 and HR2.....	60
Figure 4.7. Example of scores for item # 3 by case number for HR1 and HR2.....	60
Figure 4.8. Example of scores for item # 4 by case number for HR1 and HR2.....	61
Figure 4.9. Example of line breaks removed from response and replaced with a space.....	63
Figure 4.10. Agreement Measures and Reliability Coefficients for Item #1.....	72
Figure 4.11. Score Discrepancy for Item #1.....	74
Figure 4.12. Agreement Measures and Reliability Coefficients for Item #2.....	76
Figure 4.13. Score Discrepancy for Item #2.....	78
Figure 4.14. Agreement Measures and Reliability Coefficients for Item #3.....	80
Figure 4.15. Score Discrepancy for Item #3.....	82
Figure 4.16. Agreement Measures and Reliability Coefficients for Item #4.....	84
Figure 4.17. Score Discrepancy for Item #4.....	86
Figure 4.18. Agreement Measures and Reliability Coefficients for Item #1—Application of AES Model.....	91
Figure 4.19. Score Discrepancy for Item #1—Application of AES Model.....	93
Figure 4.20. Agreement Measures and Reliability Coefficients for Item #2—Application of AES Model.....	95

List of Figures (cont'd)

Figure 4.21. Score Discrepancy for Item #2—Application of AES Model.....	96
Figure 4.22. Agreement Measures and Reliability Coefficients for Item #3—Application of AES Model.....	.98
Figure 4.23. Score Discrepancy for Item #3—Application of AES Model.....	99
Figure 4.24. Agreement Measures and Reliability Coefficients for Item #4—Application of AES Model.....	101
Figure 4.25. Score Discrepancy for Item #4—Application of AES Model.....	102

Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education

Chapter 1: Introduction

Traditional education in healthcare has focused heavily on foundational knowledge such as anatomy and pathophysiology. While this knowledge is clearly important, the application of this knowledge to reason through a clinical problem is perhaps equally, if not more important. Breakdowns in this process can result in clinical error and patient harm (Norman & Eva, 2010). It is clearly important that health care providers develop critical thinking skills during their education. However, the challenge is how to best teach and assess these skills (Bowen, 2006; Fleming, Cutrer, Reimschisel, & Gigante, 2012).

The Canadian Patient Safety Institute (2008) recognizes the importance of accurate and comprehensive assessment of health care provider performance to ensure safe patient care and competent clinical practice. Assessment of knowledge acquisition can often be accomplished through the use of selected-response assessments (such as multiple-choice questions). These areas of knowledge acquisition that are more conducive to assessment with selected-response assessments include: anatomy, physiology, pathophysiology, pharmacology, growth and development, and patient care protocols. Specific areas in health sciences education that are difficult to accurately assess with selected-response item format examinations include: critical thinking, clinical reasoning, and clinical judgement (Fleming, Cutrer, Reimschisel, & Gigante, 2012; Kellogg & Raulerson, 2007; Minnich et al, 2018; Oermann & Gaberson, 2017). Andrich and Marais (2018), Foltz (2016), and Harmes, Welsh, and Winkelman (2016) acknowledged that written assessments are one of the most powerful methods for assessing higher-level thinking skills such as critical thinking, communication, problem solving, and creativity. Knowledge

acquisition in these areas is more accurately assessed with constructed-response item format examinations such as essays and short-answer item formats. The importance of using open-ended writing assessments to facilitate higher-level thinking, including making connections and critical thinking, has been shown in several studies (Andrich & Marais, 2018; Kellogg & Raulerson, 2007) and has resulted in an increased interest in the use of essay-type responses in examinations. Constructed-response types of assessments often include lengthy written responses which are challenging to score in relation to cost, time, and human rater reliability (Andrich & Marais, 2018; Saville, 2012; Williamson, Xi, & Breyer, 2012).

The challenges with scoring written assessments have resulted in a demand for the development of efficient assessment methods that can assess higher-level thinking skills of students (Harmes, Welsh, & Winkelman, 2016; Shermis & Burstein, 2013; Yang, Liu, & Morell, 2018). One potential strategy to address the challenges of scoring essay-type responses is automated essay scoring (AES). The increase in preference to include essay-type responses in education fields has led to an increased interest in automated essay scoring (Haberman & Sinharay, 2010; Reilly, Stafford, Williams, & Corliss, 2014; Williamson, Xi, & Breyer, 2012; Yang, Liu, & Morell, 2018).

Automated essay scoring (AES) is a developing technology that is increasingly recognized as a potential strategy for managing the challenges associated with testing and scoring written assessments, particularly in large classroom sizes (Dikli, 2006; Gierl, Latifi, Lai, Boulais, & De Champlain, 2014; Reilly, Stafford, Williams, & Corliss, 2014; Yang, Liu, & Morrell, 2018). Automated essay scoring uses the foundations of computational linguistics and computing science to build computer models to score constructed-responses (Albon, 2018; Brew & Leacock, 2013). The AES computer software incorporates natural language processing and

machine learning methods to learn from a set of previously scored essays to build a model which is used to score essays (Albon, 2018; Bennett & Zhang, 2016).

A familiar example of natural language processing and machine learning is the processing techniques that are used to autocorrect what we key into our smartphones or personal computers. The software used in these programs recognizes and suggests corrections for words that typically follow in a sentence. Processing methods such as autofill and autocorrect are techniques that most of us have become familiar with. The program learns from previous examples what words or information may be needed. Also, and possibly more importantly, the software processing adjusts to the user's responses. For example, if a name is included in a message, the software may identify it as incorrect and possibly autocorrect it. A specific example is the use of the name *Taylir*. The software will autocorrect this to *Taylor* and identify that *Taylir* was misspelled. However, as the user continues to include the name *Taylir* in responses, the program learns that this is the correct spelling for this particular user and incorporates this into the program. Therefore, the spelling *Taylir* is no longer autocorrected. The program uses previous examples to learn the correct responses.

Automated essay scoring systems use previously scored essays to teach the computer program how to score new essays. Multiple scored essays are input to the program to train the computer software. This is called the training set and requires large amounts of scored responses to effectively teach the program. Once the program has learned what the expectations for the assessment are, the program then applies this to score a new set of essays. The actual science underlying the development of natural language processing and machine learning is beyond the scope of this study. However, it is this science that provides the foundations for automated essay

scoring software programs. The focus of this study is on the application of AES with acknowledgement of the processing methods that are foundational to AES.

Automated essay scoring has many potential advantages for scoring essay-type responses including reduced time and costs, provision of immediate feedback, and improved quality of scoring (Foltz, 2016; Gierl, 2014; Kuo, Chen, Yang, & Mok, 2016; Williamson, Xi, & Breyer, 2013). It is because of these advantages that AES can potentially become a more integral part of scoring written assessments in nursing education.

Background of the Issue and Target Audience

Ensuring patient safety is the foundational guiding principle for all health sciences education (Canadian Patient Safety Institute, 2008). The concepts of patient safety are incorporated into all aspects of nursing education to ensure this standard is met. Nursing education programs across Canada include fundamental knowledge areas such as anatomy, physiology, pathophysiology, and pharmacology. These areas of knowledge are typically assessed with selected-response style items. What is often considered more critical in nursing education is the application of foundational knowledge to patient situations. This application requires the development of critical thinking skills, clinical reasoning ability, and clinical judgement skills (Brookhart & Nikito, 2015; Scriven & Paul, 2013). Breakdowns in these processes can result in clinical error and patient harm (Norman & Eva, 2010). It is clearly important that health care providers develop critical thinking skills during their education to ensure the standard of patient safety is met.

The importance of accurate and comprehensive assessment of health care providers to ensure safe patient care and competent clinical practice is paramount (Canadian Patient Safety Institute, 2008). It is widely recognized that there are specific areas in health sciences education

that are difficult to accurately assess with selected-response item format examinations including critical thinking, clinical reasoning, and clinical judgement (Bowen, 2006; Fleming, Cutrer, Reimschisel, & Gigante, 2012; Kellogg & Raulerson, 2007; Oermann & Gaberson, 2017). In order to accurately and comprehensively assess these higher-level thinking skills, written assessments are recognized as an effective assessment (Andrich & Marais, 2018; Johnson, Schwartz, Lineberry, Rehman, & Soo Park, 2018; Tankersley, 2007). However, the inclusion of short-answer essay type questions, writing assessments, and constructed-response items is, in fact, being eliminated in nursing education programs. Nursing educators across Canada and USA were contacted to provide examples of constructed-response items for this study. The author of the current study contacted educators from the Faculty of Nursing at the following programs: University of Alberta, MacEwan University, Athabasca University, Windsor University (Windsor, Ontario), University of Saskatchewan, Camosun College (Victoria, British Columbia), Lander University (Greenwood, South Carolina, USA), Eastern Kentucky University (Kentucky, USA), Gwynedd Mercy University (USA), University of Massachusetts (USA), Indiana University Northwest (USA), Widener University (USA), and Gadsden State Community College (USA). Not one of the programs include short-answer essay type items in any of their classes that have class sizes larger than 20. Many reported that they haven't included short-answer essay type questions for the past 10 or more years.

When the author was unable to locate any examples of short-answer essay type questions in the contacted nursing education programs, educators in other related areas of health sciences were contacted. These include: Faculty of Pharmacy and Pharmaceutical Sciences at the University of Alberta, Faculty of Rehabilitation Medicine at the University of Alberta, Department of Health Sciences at NorQuest College, Department of Physiology, Athabasca

University. Again, not one of these programs include short-answer essay type items in their assessments in class sizes greater than 20. Finally, the author contacted Elsevier Canada and Wolters Kluwer Health/Lippincott, Williams, & Wilkins publishing companies to request any examples of short-answer essay type items. None of these contacted publishers include short-answer essay type items in their assessments.

All the educators contacted identified that they would like to include short-answer essay type items in their assessments and, also, noted that they agreed these assessments were the most accurate in assessing higher-level thinking skills in nursing students. However, these educators noted the same challenges that have resulted in excluding short-answer essay type items in their assessments—too costly and too time-consuming to score. A few of the educators also included lack of interrater reliability as an additional challenge. Given these results, how are higher-level thinking skills being assessed to ensure the standards of safe patient care are met in nursing education programs?

Another interesting point is that every educator that was contacted reported that they would like to include short-answer essay type items in their examinations and assessments. All of these educators requested to be contacted “if AES works” so that they can include these types of assessments in their courses again.

Automated essay scoring has been available for decades and it is a possible solution to overcome the challenges in scoring assessments of higher-level thinking skills in nursing education. However, the information and awareness of AES systems is lacking. There are no examples of AES being used in nursing education in the literature. At this point in time, it appears that AES has not been used to score any assessments in nursing education. This lack of information and awareness results in a decreased acceptance of AES as a potential strategy to

overcome the challenges associated with scoring short-answer essay type items. Also, Bennett and Zhang (2016) recently noted that the challenges in explaining and defending AES technology results may also cause a lack of acceptance and use of AES in all education fields.

Another important factor is the resistance to using computers to score assessments rather than humans. Historically, humans have resisted the utilization of machines (including computers) to perform any tasks that humans were performing until there was evidence that machines (including computers) could do the task as well as, or better than, humans (Shermis, 2014). For example, even simple parking pay machines and automated call answering systems are typically met with frustration by humans until it is demonstrated that the machines are at least as effective as humans, if not more effective. The implementation of advanced machines to dispense medications, control intravenous infusion rates, and monitor patient hemodynamics has been met with resistance from nurses and other health care professionals over the past several decades (Barrett & Stephens, 2017; Byers, 2017). This may also mean that there may be resistance among nursing educators for acceptance of computers scoring essays rather than humans. Educators may question how a computer program could possibly read and accurately score a written response to a higher-level thinking patient care question. Therefore, more research is needed in this area to explore the effectiveness of AES for scoring assessments in nursing education.

Focus of the Research Study and Research Questions

The focus of this dissertation research is to evaluate the effectiveness of using AES for scoring short constructed-response type items based on a patient scenario in nursing education to assess higher-level thinking skills. It is suggested that using only selected response style items for assessment of learning is inadequate to assess critical thinking, clinical reasoning, and

clinical judgement related to patient scenarios (Fleming, Cutrer, Reimschisel, & Gigante, 2012; Kellogg & Raulerson, 2007; Oermann & Gaberson, 2017). The incorporation of constructed-response style items in examinations and assessments for nursing students has been met with resistance because of time and cost to score these items. Evaluating the effectiveness of AES for scoring constructed-response items related to patient scenarios may give us information to help improve the assessment of learning in the areas of critical thinking, clinical reasoning, and clinical judgement. Accurate assessment of these essential components of nursing practice impacts the safety of patients, clients, families, and populations.

Key questions related to this proposed research are:

- 1) Is AES as effective as human raters for scoring constructed-response items in nursing education in terms of accuracy and reliability?
- 2) Does AES score constructed-response items more efficiently than human raters?
- 3) Is AES a potential solution to overcome the challenges of time, cost, and subjectivity in scoring constructed-response items?

Importance of this Research Study

It is well recognized that accurate and comprehensive assessment of higher-level thinking skills is essential in nursing education to ensure the standards of safe patient care (Canadian Patient Safety Institute, 2008; Oermann & Gaberson, 2017). Critical thinking skills, clinical reasoning ability, and clinical judgement skills have been identified as higher-level thinking skills that are essential for nursing practice to ensure safe patient care (Alfaro-LeFevre, 2017; Gordon, 2000; Oermann & Gaberson, 2017; Posel, McGee, & Fleiszer, 2014; Raymond-Seniuk & Profetto-McGrath, 2011; Scriven & Paul, 2013) and that breakdowns in these processes can result in clinical error and patient harm (Norman & Eva, 2010). It is important that health care

providers develop critical thinking skills during their education to ensure the standard of patient safety is met. But there are significant challenges to assessing the development and application of these skills. By eliminating constructed-response type items from assessments in nursing education, it is difficult to accurately assess whether nursing students are meeting the standards for safe patient care. It is essential that we find an effective solution to overcome the challenges impacting the reduction in use of short-answer type items in assessments in nursing education. This study will give more information into the potential use of AES in nursing education.

Dissertation

This dissertation is organized into five chapters. These chapters are structured as follows: 1) chapter one (the current chapter) includes introductory information, a brief overview of AES and the literature, background information related to the research issue, and a description of the research questions and study; 2) chapter two includes a review of selected literature on critical thinking in nursing practice, clinical reasoning and clinical judgement in nursing practice, critical thinking and patient scenarios, constructed and selected-response item formats, AES, feedback and learning, and the potential for AES in scoring, providing feedback, and assessing student performance in nursing education; 3) chapter three includes a detailed description of the study design, methods, data collection, AES system, and data analysis techniques; 4) chapter four includes the results and data analysis as well as the interpretation and discussion of the findings; 5) chapter five includes the summary of findings and conclusions of the study as well as a discussion of possible limitations and future directions for research.

Chapter 2: Literature Review

The following is a review of relevant research studies and literature in the following areas: critical thinking; critical thinking in nursing practice; clinical reasoning and clinical judgement in nursing practice; critical thinking and patient scenarios to assess learning; feedback and learning; selected-response and constructed-response item formats; automated essay scoring; and the relevance of these areas in nursing education. Specifically, the focus of this literature review is on assessing higher-level thinking skills such as critical thinking, clinical reasoning, and clinical judgement in nursing education and commonly used learning assessments.

The literature review is organized into two sections. The first section overviews higher-level thinking skills in nursing education such as critical thinking, clinical reasoning, and clinical judgement. Section one also includes an overview of commonly used learning assessments in nursing education such as patient scenarios, selected-response items, constructed-response items, and the importance of feedback in learning. Section two is an overview of automated essay scoring including challenges, advantages and disadvantages, and a description of AES systems that may be useful for learning assessments in nursing education. The purpose of this literature review is to outline the issues affecting learning assessments in nursing education and explore AES as a potential solution to overcoming these challenges in nursing education.

Section One: Assessing Higher-Level Thinking Skills in Nursing Education

One of the most challenging aspects of nursing practice includes the ability to think critically, clinically reason, and make sound judgements to guide effective patient care. It is well documented that effective clinical reasoning skills in nursing practice are essential to safe and effective patient care (Kuiper & Pesut, 2004; Kuiper, Pesut, & Kautz, 2009; Murphy, 2004) and Turkel and Morrison (2016) noted that underdeveloped critical thinking skills in nurses

compromises patient safety.

Critical Thinking

There are many definitions of critical thinking (Zori, 2016), but it is commonly recognized that critical thinking is a set of skills and behaviours that guide thought processes for decision making and actions (Alfaro-LeFevre, 2017; American Philosophical Association, 1990; Raymond-Seniuk & Profetto-McGrath, 2011; Zori & Morrison, 2009). The skills of interpretation, analysis, evaluation, inference, explanation, and self-regulation together with the dispositions of systematicity, analyticity, inquisitiveness, truth seeking, self-confidence, open-mindedness, and maturity are often included in definitions of critical thinking (Alfaro-LeFevre, 2017; American Philosophical Association, 1990; Edwards, 2007; O’Neill & Dluhy, 1997).

Scriven and Paul (2013) defined critical thinking as the “intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication” (p. 1). Other authors have commented on the importance of reflection and metacognition on thinking critically (Alfaro-LeFevre, 2017; Raymond-Seniuk & Profetto-McGrath, 2011). Successful critical thinking requires the availability of necessary domain knowledge, association of this knowledge with evidence-based research and then subsequent application of this knowledge through decision making, high-quality clinical judgements, and problem solving (Gordon, 2000; Posel, McGee, & Fleiszer, 2014).

Critical thinking is an important aspect in most fields of knowledge and practice, but it is recognized as particularly important in nursing practice because of the potential impact on patients (Alfaro-LeFevre, 2017; Zuriguel Perez, Lluch Canut, Falco Pergueroles, Puig Llobet, Moreno Arroyo, & Roldan Merino, 2014).

Critical Thinking in Nursing Practice and Dual Process Theory

Critical thinking is considered essential for nursing education and nursing practice (Alfaro-LeFevre, 2017; American Association of Colleges of Nursing, 2008; Cazzell & Anderson, 2016; Kuiper, Pesut, & Kautz, 2009; Raymond-Seniuk & Profetto-McGrath, 2011; Zuriguel Perez et al, 2014). The complexity and constantly changing nature of the health care environment combined with the focus on evidence-informed patient care highlights the importance of critical thinking in nursing practice (Dexter et al, 1997; Zuriguel Perez et al, 2014).

Critical thinking is described as cognitive processes that analyze information to facilitate clinical reasoning, judgement, and decision-making (Alfaro-LeFevre, 2017; Cazzell & Anderson, 2016; Dexter et al, 1997; Zuriguel Perez et al, 2014). There have been several studies conducted to better understand how clinicians make decisions. Although several frameworks have been proposed, one of the most commonly cited is the dual process theory (Croskerry, 2009; Kahneman, 2011; Pelaccia, Tardif, Tribby, & Charlin, 2011). This framework has its origins in the cognitive psychology literature and describes two cognitive processes responsible for decision making: 1) a rapid intuitive system (System1); and 2) a system of more deliberate analysis (System 2). Although the dual processing theory was developed outside of health care, it has been used as a framework to often explain critical thinking in health care. Balla and colleagues (2012) interviewed health care providers following patient encounters. The majority of health care providers developed a hypothesis about the cause of the patient's health issue early in the encounter (System 1) and then actively collected data to more critically analyze that initial intuition (System 2) (Eva, 2005; Pelaccia et al., 2011). By collecting and critically analyzing

patient data, clinicians can reason through the initial patient assessments and make decisions and judgements about effective intervention plans for the patient.

Unfortunately, there are challenges in teaching and assessing these clinical reasoning and decision-making skills in clinicians. Typically, it is assumed that these skills are developed concurrently with domain-specific knowledge (Kassirer, 1995), though learners' skills in critical thinking are often found to be poor. In one study of graduating nurses, it was demonstrated that only 35% met expectations of critical judgement as measured by a standardized validated metric (del Bueno, 2005). Cazzell and Anderson (2016) noted the gap between nursing education and nursing practice related to critical thinking and clinical judgement needs to be addressed and that more research into the development and assessment of critical thinking in nursing education is needed.

One strategy that has been suggested to assess learner's critical thinking skills is to question them about their initial impression of a patient early in the data acquisition process (Bowen, 2006; Rencic, 2011). Traditionally, learners are asked to present details about a patient after they have already performed a detailed history and physical examination. By asking students early in the data acquisition process, they are forced to develop early hypotheses based on the available data (System 1). These early hypotheses can then help focus the subsequent history taking and physical examination to prove (or disprove) the working hypothesis (System 2).

This is essentially a think-aloud process in which learners are required to verbalize their cognitive processes as they work through a case (Bowen & Ilgen, 2014; Burbach, Barnason, & Thompson, 2015). Part of this process is to evaluate how learners are synthesizing the data they are obtaining. Semantic qualifiers refer to the meaning that health care providers attach to the

clinical data (Bowen, 2006). For example, if a patient is reporting joint pain that started last night and is similar to pain experienced several years ago, then this information is classified as acute, recurrent pain which immediately narrows the list of diagnostic possibilities. Clinicians who use more semantic strategies when discussing patient clinical situations are typically more likely to arrive at the correct diagnosis (Chang, Bordage, & Connell, 1998).

Critical Thinking Theoretical Framework: A Two-Phase Framework

In addition to the dual process theoretical framework, another foundational framework for critical thinking in this proposed research study is the *Two-Phase Critical Thinking Framework* by Edwards (2007). Edwards (2007) outlined two phases of learning and integration of critical thinking skills in nursing students and developed the framework to guide teaching and assessment of critical thinking skills in nursing education.

Phase one of this framework includes the following components: interpretation and organization of the information; hidden assumptions; breaking down the situation; consideration of all the options; subjective and objective knowledge; conflicting issues; and decision-making. The first phase of this framework for critical thinking in nursing practice involves gathering and analyzing all available information about the situation and integrating this information with knowledge of potential solutions then making decisions to guide patient care (Edwards, 2007). This phase is about considering and selecting alternatives and planning which actions and direction to take (Edwards, 2007). As outlined in the literature on critical thinking, it is essential to use critical thinking skills to analyze, evaluate, interpret, and explain information for clinical reasoning and decision-making to guide safe patient care (Alfaro-LeFevre, 2017; Edwards, 2007; Raymond-Seniuk & Profetto-McGrath, 2011, Zori, 2016). It is essentially the combination of critical thinking skills with knowledge to reason through a clinical situation to formulate

decisions to guide safe patient care and clinical practice (Edwards, 2007; Raymond-Seniuk & Profetto-McGrath, 2011). Once the decisions are made, the learner then proceeds through phase two.

After the learner has progressed through phase one and decided on the direction for patient care, the learner then proceeds to phase two where the process of the decision-making is considered and evaluated. Phase two of the framework includes the explanation of the decision, accountability and responsibility for the decision, evaluation of the process, and creativity and innovation to move forward and learn (Edwards, 2007). Phase two is about justifying and being accountable for the decisions then reflecting on the decision-making process and considering new initiatives or policies to guide future practice (Edwards, 2007).

Phase one and two are dynamic phases and the learners move back and forth between phases to reconsider, evaluate, and reflect on the processes involved in clinical decision-making. A pictorial explanation is included below (Fig. 1) which emphasizes both phases and how the framework guides critical thinking.

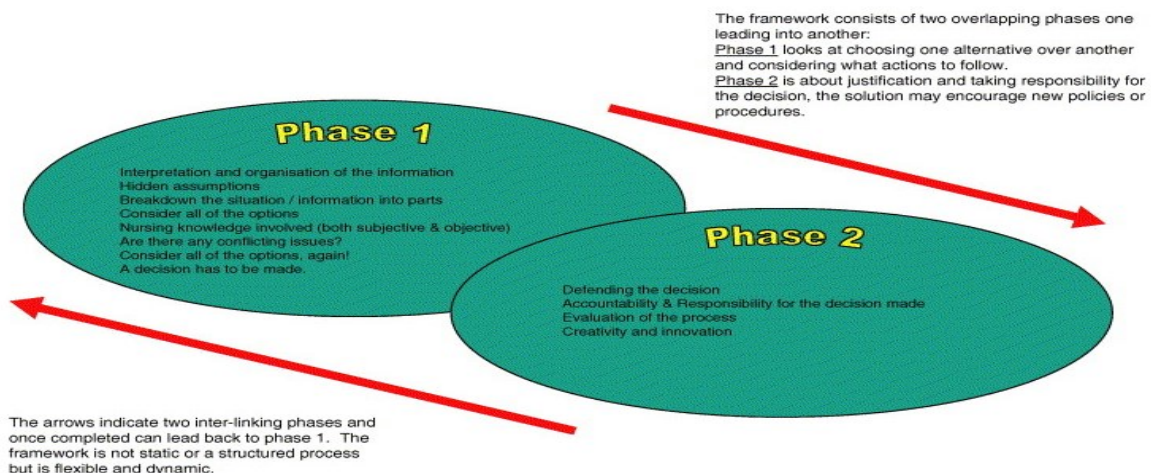


Figure 1: Critical Thinking: A Two-Phase Framework. Used with permission; Edwards, S.L. (2007). Critical thinking: A two-phase framework, *Nursing Education in Practice*, 7, 303-314.

Critical Thinking: A Two-Phase Framework Applied to a Patient Scenario

Consider the following patient scenario: Mrs. S., a 36 year old patient, reports that she has a “really bad headache”. She informs you that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

The critical thinking two-phase framework can be applied to this patient scenario to exemplify how clinicians critically think to reason and make decisions. In phase one, the clinician would gather information about the patient and her headache, including a symptom analysis (severity, associated factors, alleviating or aggravating factors, treatments, timing, quality of pain), existing and previous health conditions, medications, patient allergies, known risk factors, and other relevant information. The clinician would then integrate all this subjective and objective information with knowledge and consider other assumptions such as migraine and stroke risk for this age and gender, possibility of drug use, mental health issues, and others. Then the clinician considers all possible options such as dehydration, migraine, stroke, aneurysm, or allergic reactions then analyzes all the available information and knowledge to make decisions about the direction of care for this patient. This analysis and clinical reasoning for decision-making is crucial for the outcomes for this patient. If the clinician makes the decision that the patient may be dehydrated, different diagnostic tests are considered and the patient can consume fluids. If the clinician makes the decision that the patient may be experiencing a stroke, then a very different path of care will be initiated for the patient.

Phase two of this framework applied to this scenario would include following through on the decision-making process, reflecting and evaluating the success of the decision-making process, and possibly initiating different practice policies and guidelines for future care for

similar patient situations. The clinician in this example would evaluate the process, justify the decisions made, and take responsibility for correct and incorrect decisions.

Critical Thinking: A Two-Phase Framework Applied to this Research Study

The Two-Phase Framework for Critical Thinking by Edwards (2007) can also be used as the framework for this overall research study. In this project, phase one would reflect the processes of gathering information, hidden assumptions, analysis of the situation, consideration of the options, subjective and objective knowledge, conflicting issues, and decision-making all in relation to how higher-level thinking skills are assessed in nursing students. Phase one would include all the information gathering about the importance of including written assessments for assessing critical thinking, clinical reasoning, and clinical judgement in nursing students. Also included would be the assumptions that scoring written assessments is not possible in large class sizes, issues with time and cost of these assessments, and further analysis of how higher-level thinking skills are assessed in nursing education programs. By gathering and analyzing all this information, potential solutions and alternatives can be identified and planned to inform the direction to take.

Phase two, as applied to this overall study, involves explanation of decisions and implementing the strategies identified. Also included in phase two is the accountability and responsibility for the direction taken and reflection on the processes to determine future steps and initiatives. In this research study, the decision to include written assessments in examinations followed by the implementation of this decision required several steps. These steps included developing and reviewing the items, structuring the computer-based format and platform for the items, and then inclusion of the items on computer-based examinations administered through *eclass*.

As the results and analysis of this study become available, this information can be used to guide future practice in assessing higher-level thinking skills in nursing students.

Clinical Reasoning and Clinical Judgement in Nursing Practice

Some of the most challenging aspects of nursing practice include the ability to clinically reason, think critically, and make sound judgements to guide effective patient care. The importance of well-developed clinical reasoning skills in nursing practice is essential to the safe and effective delivery of patient care (Kuiper & Pesut, 2004; Kuiper, Pesut, & Kautz, 2009; Murphy, 2004). Nurses with effective clinical reasoning skills will most likely have a positive influence on patient outcomes and, conversely, nurses with ineffective clinical reasoning skills and who miss critical information from patients will often result in adverse patient outcomes (Aiken et al, 2003; Lapkin, Levett-Jones, Bellchambers, & Fernandez, 2010; Shellenbarger & Robb, 2015). Koharchik, Culleiton, Caputi, & Robb (2015) describe clinical reasoning as essential for “preserving the standards of the nursing profession and promoting good patient outcomes” and identified the development of clinical reasoning skills as the “crux of nursing education” (p. 58). Often the terms clinical reasoning and clinical judgement are used synonymously. However, it is recognized that clinical reasoning is broader and includes the processes that lead to and facilitate sound clinical judgements which promote effective patient care (Harmon & Thompson, 2015; Koharchik, et al, 2015; Shellenbarger & Robb, 2015).

Shellenbarger and Robb (2015) note that clinical reasoning involves the collection of information from the patient, processing the information collected, planning and implementing appropriate interventions, and evaluating the clinical reasoning process. The knowledge and ability to assess a clinical situation, identify interventions, and plan and implement an appropriate course of action are the essence of clinical judgement (Edwards, 2007; Huckaby,

2009; Lindsey & Jenkins, 2013; Simmons, 2010; Tanner, 2006). Clinical reasoning, judgement, and critical thinking skills increase with experience (Lindsay & Jenkins, 2013). Several teaching methods and learning experiences, such as concept-mapping, patient scenarios, simulation, and reflective journaling, have been used to foster the development of these skills in nursing students (Cappelletti, Engel, & Prentice, 2014). One of the continuing challenges is how to then assess these skills in nursing students. Observation and evaluation of students' performances in clinical practice and simulated experiences are valuable to assess critical thinking, clinical judgement, and clinical decision making skills in students. Short essay questions or written papers are other valuable assessment tools to assess these skills in nursing students because the complexity of patient situations is more accurately assessed with constructed-response type items.

Constructed-response item formats are considered to be more authentic in assessing performance and can provide more real information about the performance of the examinee (Brookhart, 1993; Gierl et al, 2014; Ramineni & Williamson, 2013).

Critical Thinking, Clinical Reasoning, and Clinical Judgement in Nursing Education

The importance of critical thinking, clinical reasoning, and clinical judgement in nursing practice is well recognized and documented. One of the challenges in nursing education is how to teach and assess these higher-level thinking skills (Dexter, 1997). Several strategies to facilitate learning higher-level thinking skills are implemented in nursing education and some examples include: context-based learning with patient scenarios, concept-mapping, simulation, virtual simulation, and reflective journaling.

Context-based learning with patient scenarios. In context-based learning, similar to problem-based learning, learners are given a situation (usually based upon real life situations) and the learners work through the situation to assess the issues, gather information, plan

interventions, implement interventions, and evaluate the processes they used to determine the effectiveness of their strategies (Zuriguel Perez et al, 2014). Learners draw upon foundational knowledge such as anatomy, physiology, and pharmacology, to assess, identify, plan, and reason through the situation. Learners work individually or in teams to resolve the situation. Carter, Creedy, and Sidebotham (2016) identified problem-based learning as one of the most commonly used strategies to learn critical thinking skills.

Concept-mapping. Concept-mapping is another common strategy used in nursing education to teach critical thinking skills (Carter, Creedy, & Sidebotham, 2016; Cazzell & Anderson, 2016; Zori, 2016). Learners draw out diagrams that identify and highlight the relevant concepts in a particular situation and connect related ideas and concepts to describe the relevant issues. Definitions and relationships between concepts are included in the maps to give an overall schematic of the situation and potential solutions.

Simulation. Simulation in health care education is a pedagogical approach that realistically replicates clinical situations for the purposes of learning and practice for health care professionals. Human patient simulation is increasingly used in health care education to promote patient safety, critical thinking skills, interprofessional teamwork, leadership, and professional competencies, and is widely accepted as an effective learning methodology to achieve these outcomes (Conrad, Guhde, Brown, Chronister, & Ross-Alaolmolki, 2011; Curtin, et al. 2011; Dillon, Noble, & Kaplan, 2009; Greidanus, King, LoVerso, & Ansell, 2013; Hall, Soderstrom, Ahlqvist, & Nilsson, 2011; Lapkin, Fernandez, Levett-Jones, & Bellchambers, 2010). An important aspect of simulated learning is that learners can practice and learn in a realistic clinical situation and safely make mistakes without harming an actual patient. Simulated experiences include the use of actors who pretend to be patients, role playing, case studies, virtual

simulations (similar to gaming), and incorporating computerized high fidelity mannequins that breathe, blink, and respond to interventions similarly to human responses. The mannequins are situated in clinical rooms that replicate the clinical setting and the goal is to make the clinical simulation as close to reality as possible.

Virtual Simulation. An increasingly used technique in simulation to teach and assess critical thinking and clinical reasoning is screen-based simulations or virtual patients (Posel et al., 2014). These simulations are “a specific type of computer program that simulates real-life clinical scenarios; learners emulate the roles of health care providers to obtain a history, conduct a physical exam, and make diagnostic and therapeutic decisions” (Candler, 2007). The learner is presented with a “patient” on a computer screen including vital signs, laboratory results and relevant multimedia such as electrocardiograms and video. They are then required to select from a list of optional steps to best diagnose and manage the patient (such as applying oxygen by face mask or starting cardiopulmonary resuscitation (CPR)). These simulation systems allow educators to tailor-make scenarios that best highlight specific critical thinking skills. Learners can practice scenarios repeatedly without an instructor present.

Reflective Journaling. Another strategy used in nursing education to foster critical thinking skills is the use of reflective journaling and writing. Learners write about situations they have experienced, topics they are interested in, challenges they have been exposed to, and successes and failures in their lived experiences. These writing exercises facilitate learners to reflect and consider how these situations and experiences have impacted them and their learning. Reflective journaling requires learners to take an in-depth look into issues and experiences and to think about the experiences and influences on themselves.

Critical Thinking and Patient Scenarios to Assess Learning

Several different methods have been proposed to assess critical thinking of health sciences students, including the use of patient scenarios. Students are given a scenario about a patient, which includes some basic information and possible health assessment data. Students are then asked to respond to questions about the patient scenario to develop appropriate plans of care for the patient. Specific areas of questioning include: what additional information should be collected; physical examinations; possible diagnostic tests; appropriate interventions; and evaluation of the plan. Learners emulate the roles of health care providers to obtain a history, conduct a physical exam, and make diagnostic and therapeutic decisions (Candler, 2007; Oermann & Gaberson, 2017). These scenarios allow educators to tailor-make scenarios that best highlight specific critical thinking skills. However, it is important that learners receive feedback on their performance.

In order to receive feedback to improve the effectiveness of learning, it has been suggested that students could enter some longer form responses that highlight their cognitive processes as they work through the case (Brookhart & Nikito, 2015; Oermann & Gaberson, 2017; Posel et al., 2014). This information can then be submitted to a facilitator for feedback. This method is time consuming to score, facilitator-dependent and does not provide immediate feedback to the students. In instances where class sizes are larger, it could take weeks for an educator to read, score, and provide feedback to all the learners. Even with a class size of 20 students it is challenging to score written assessments and provide feedback to the learners in a timely matter. This is an important challenge affecting the use of written responses for assessment of learning.

Feedback and Learning

There are multiple definitions of feedback. Generally, feedback is considered to be an interactive process that provides learners with information about their performance which then helps the learner improve future performance. It is widely acknowledged that feedback is a critical element of learning (Clynes & Raftery, 2008; Cushing, et al, 2011; Goodman, Wood, & Chen, 2011; Lam, DeRue, Karam, & Hollenbeck, 2011; Schlegel, Woermann, Rethans, & van der Vleuten, 2012). It is also well recognized that a critical aspect for effective learning in health sciences education is guided feedback (Alinier, 2010; Burke & Mancusco, 2012; Issenberg, McGaghie, Petrusa, Gordon, & Scalese, 2005; Motola, Devine, Chung, Sullivan, & Issenberg, 2013). Guided feedback is the process of providing information to the learners about their performance and also includes helping the learners to recognize and develop feedback for their own individual performances (Alinier, 2010; Burke & Mancusco, 2012; Motola, Devine, Chung, Sullivan, & Issenberg, 2013).

Feedback can be described along several dimensions such as specificity, frequency, timing, source, and type (Lam, et al. 2011). In addition to these, Shute (2010) discussed the dimensions of verification, elaboration, complexity, and length of feedback. These dimensions are relevant to how feedback is provided to learners.

Specificity refers to the level of detail of feedback given to the learner. Feedback can range from a general, broad statement about the learner's overall performance to a detailed review of a particular aspect of the learner's performance. Goodman, Wood, and Chen (2011) hypothesized that highly specific feedback can be prescriptive and result in learners receiving high levels of guidance. Less specific feedback provides less guidance and may help learners manage more challenging aspects of a task. Generally, it appears that different learners and

learning outcomes may have different needs for the level of specificity of feedback (Shute, 2010).

Frequency of feedback refers to how often the learner receives information regarding their performance. Research has shown differing results about the frequency of feedback that is most effective for learning (Lam, et al. 2011; Parekh & Thorpe, 2012).

Timing refers to the immediacy of receiving feedback. Feedback can be provided throughout a particular experience or can be given at the end of a session. Several studies have looked at the effect of timing of feedback on the effectiveness of learning (Clynes & Raftery, 2008; Corrigan & Hadham, 2011; Cushing, et al. 2011) and have noted that feedback received closer to the assessment is more effective for learning than feedback given at a later time.

Source of the feedback refers to who is giving the feedback. Research into feedback given by instructors (Goodman, et al. 2011; Lam et al, 2011; Oestergaard, et al. 2012), peers (Cushing, et al. 2011; Stegman, et al. 2012), and self-reflection (Alinier, 2010; Corrigan & Hardham, 2011; Motola, et al, 2013) demonstrates the important role that feedback has on learning and that there can be multiple sources of feedback for learners. This is an important factor to note because learners are generally open to receiving feedback from sources other than their instructors, including feedback from a computer (Walkow & Reilly, 2014).

It is well recognized that a critical aspect for effective learning in nursing education is the feedback that is provided for the learners (Alinier, 2010; Burke & Mancusco, 2012; Issenberg, McGaghie, Petrusa, Gordon, & Scalese, 2005; Motola, Devine, Chung, Sullivan, & Issenberg, 2013). The elements of timing, specificity, frequency, and source of feedback are all important to consider when assessing nursing student performance. Giving specific, individualized feedback in a timely manner is beneficial to effective learning. However, it may be time

consuming, costly, and challenging to accomplish these outcomes. Instructors may forget which aspects of the examination were completed or performed accurately and with large number of learners, many different facilitators are required resulting in challenges in discrepancies in the feedback and teaching provided. Also, learners may want to receive feedback in a more private and individualized manner. In order to achieve this level of private and individualized feedback, educators would need to schedule one-to-one discussions with the learners to review and discuss the learner's performance. This would be very time-consuming.

Selected-Response and Constructed-Response Item Formats

There are essentially two major types of response item formats that are used in assessing the performance of learners. Selected- and constructed-response item formats are predominantly used in education and have several advantages and disadvantages. Some assessment methods used in educational examinations include one of the item formats used alone or both formats used together in a mixed format assessment.

Selected-Response Item Formats. One of the major challenges in medium and large-scale examinations is the cost of developing, administering, and scoring the exams. In order to reduce the costs associated with scoring exams, many large-scale examinations use the selected-response item format. A selected-response item format refers to the type of items where examinees choose or select from a list of options. The most commonly used format for selected-response items is the multiple-choice question. In a multiple-choice question, an examinee reads an item (stem) then selects one response from several options (distractors and key). Other selected-response formats include: extended multiple-choice (use of increased number of distractors to reduce likelihood of guessing), multiple selection (examinees select more than one correct response from the options), specifying relationships (examinees identify connections

between concepts), ordering information (examinees arrange options in a correct order), selecting and classifying (examinees categorize options with concepts), inserting text (examinees fill in the blank with correct responses), and true-false items (examinees determine if item is true or false) (Sireci & Zenisky, 2006). Although selected-response items are efficient to score, objective, measure a wide range of content and cognitive skills, and yield a high reliability, there are several disadvantages (Gierl, 2014, lecture notes). Selected-response items are costly to develop, require extensive reviews to ensure accuracy of content and options, limit types of thinking that can be examined, are subject to examinees successfully guessing, and test small amounts of content per item. In order to test large amounts of content, hundreds of questions need to be developed and reviewed. Typically, selected-response items are scored dichotomously as correct or incorrect. This dichotomous scoring of selected-response items results in decreased levels of information collected in the assessment (DeMars, 2008; Dragow, Levine, Tsien, Williams, & Mead, 1995). Finally, if the security of a selected-response item exam is breached, it is very costly to develop new items for the exam.

Constructed-Response Item Formats. Constructed-response item formats include any item that requires examinees to create their own responses or answers to the item. Examples include short-answer questions and responses, essays, stories, position papers, and scenarios. Constructed-response item formats are considered to be more “authentic” in assessing performance and can provide more “real” information about the performance of the examinee (Brookhart, 1993; Brookhart & Nikito, 2015; Gierl et al, 2014; Oermann & Gaberson, 2017; Ramineni & Williamson, 2013). Constructed-response items are easier and inexpensive to develop, have fewer challenges with security, reduce the success rate of guessing, and can be administered at different times (Yang, Liu, & Morell, 2018). However, the major challenges

associated with constructed-response item format are the costs and reliability of the scoring procedures (Kuo, Chen, Yang, & Mok, 2016). Constructed-response items require more time to score, are costly to score, can be affected by subjectivity, are susceptible to challenges with reliability (between/within examiners and essays), and are subject to scoring errors (Attali, Lewis, & Steier, 2013; Dikli, 2006; Gierl et al, 2014; Shermis & Burstein, 2013).

Section Two: Automated Essay Scoring

Automated essay scoring can be described as the use of computerized technology to score and evaluate written work (Shermis & Barrera, 2002; Shermis & Burstein, 2003) and is used to overcome the challenges of time, cost, and reliability issues in writing assessments (Attali, Lewis, & Steier, 2012; Dikli, 2006; Gierl et al, 2014). Automated essay scoring was first envisioned in 1966 by Ellis Page, who is considered the originator of AES (Gierl et al, 2014; Page, 2003). Ellis Page developed Project Essay Grader™, though many other programs for AES have been developed over the past several decades, including Intelligent Essay Assessor™ (IEA), E-rater® and Criterion™, IntelliMetric™, and MY Access!®, Bayesian Essay Test Scoring System™ (BETSY), Waikato Environment for Knowledge Analysis (WEKA) and Light Summarization Integrated Development Environment (LightSIDE). Although AES was first considered in the 1960s, it was not until the 1980s and 1990s that computer technology was available to actually implement AES in an efficient manner (McCurry, 2010).

Automated Essay Scoring System Processes

Automated essay scoring is evaluating and scoring written responses through the use of computer programs (Shermis & Burstein, 2003). Although there are several different AES systems available, they all use similar foundational processes to score written assessments. The

specific frameworks noted above have multiple unique features for each system but there are some fundamental processes similar in all AES systems such as feature extraction, machine learning and model building, and scoring. There are typically four phases in AES: 1) data preparation; 2) feature extraction; 3) machine learning and model building; and 4) scoring.

Data Preparation. The initial phase of AES is data preparation. Data preparation is not unique to AES, as all research studies require data preparation before analysis can be conducted, but the uniqueness of data preparation as it relates to AES is ensuring that the data that will be analyzed is in a format that the computer can recognize. For all AES systems, the data must be in a file format that can be input to the software program. Both the LightSIDE and WEKA systems require the data file to be in a csv format. This is a critical step in the process of using AES to score any assessments. The data must be in a format that the software program can use. If the data collected cannot be converted to a csv file, the AES system will not be able to recognize or score the data. It is important to note that csv file format is a commonly used file format and data collected within excel programs can be converted to csv files. Many nursing educators are aware that the national examination for licensure exam reports are distributed in csv file format so this is a commonly used file format in nursing education. This is an essential aspect to consider because potential users of AES must be comfortable with data preparation to use the systems.

Feature Extraction. Once data preparation is complete, the responses are input to the computer program and the AES software begins with feature extraction. Feature extraction is the process of transforming the information into a set of features (Witten, Frank, & Hall, 2013). Features are the measurable properties of the information that is being studied (Witten, Frank, &

Hall, 2013). Features are the pieces of the words that the computer program identifies and recognizes to reduce the overall amount of information to process.

Feature extraction begins from an initial set of scored data (the training set) and identifies the relevant aspects of the data. When data that is input into a computer program is large and has repetitive aspects, the process of feature extraction is used to identify and extract the key aspects (or features) that are considered to be relevant or important (Witten, Frank, & Hall, 2013). This results in a reduced data set that has the important and relevant data instead of the entire initial set of data that may include non-relevant information. Essentially, feature extraction reduces the information included in the data set to a smaller data set that predominately contains only the relevant and important information. There are multiple reduction techniques that are available to accomplish the task of feature reduction. Feature extraction requires the application of computing algorithms, cognitive sciences, and linguistic processing to identify the relevant cues in written responses (Shermis & Burnstein, 2013). Some of these include: latent semantic analysis, principal component analysis, multifactor dimensionality reduction, autoencoder, and multilinear principal component analysis. These are some examples of the computing algorithms that can be used in feature extraction and model building for AES systems.

For example, consider the following written response: “The patient reported having the most severe headache ever. The patient also reported experiencing visual changes, nausea, and numbness in her right arm.” The use of feature extraction can reduce the overall amount of information requiring processing and pull out relevant information which may then look like: “severe headache, right arm numbness, nausea, visual changes”.

Machine Learning and Model Building. Once the program has identified and extracted the features, the next step is the application of machine learning techniques to build the models

that will be used to score future responses. Machine learning is an applied computing science that allows computers to find information and learn without being explicitly programmed (SAS, 2016). Machine learning is the process of computers learning from previously input data then applying this learning to adapt to new sets of data (Alpaydin, 2014). Many examples of machine learning exist in our everyday lives such as automatically filling in online forms with correct responses to the required fields, suggestions from online shopping or search engines for consumers to consider, handwriting recognition for depositing cheques from a smartphone, and email spam filtering. Machine learning uses algorithms that learn from input data to build models to apply to new sets of information. These models can accurately and quickly analyze large, complex sets of data without any intervention from humans (SAS, 2016). More importantly, as the models are exposed to new data, the models can independently adapt to analyze the new data (SAS, 2016). This process is similar to humans building equations to solve future mathematical problems. It may take months or even years for a human to build an equation that will be used to solve future problems. A computer program can achieve the same task in minutes or even seconds. This is the same principle as machine learning. Computers can build models in seconds whereas humans may need days or weeks to build an accurate model (Witten, Frank, & Hall, 2013).

In patient care, machine learning can be used to improve efficiency and accuracy. Some examples of this include automated medication dispensary devices, laboratory analysis computer systems, and mechanical ventilation machines. Automated medication devices are computerized dispensing machines that require information to be input which the program analyzes then releases the medication for the patient based upon the input information. For example, a patient is prescribed morphine for pain management following a traumatic injury. Before the

medication is released, the computer analyzes the input information such as patient identification, medication dosage parameters, timing of last dose, patient vital signs, identification of nurse administering the medication, and concentration of morphine. This information is input to the program and analyzed by the computer. If all the input information meets the safety standards, the medication is dispensed and then administered by the nurse. Laboratory analysis computer systems analyze and compare patient blood samples and use algorithms to determine critical blood value levels. If a critical blood value level or combination of levels is detected, the computer program automatically sends alerts to the patient care unit to notify health care providers of a potentially harmful situation. Mechanical ventilation systems use computer algorithms to determine levels of oxygen, carbon dioxide, and nitrogen in a patient then adapt the mechanical ventilation systems to meet the ventilation requirements for the patient. When critical levels are detected, the program sends alerts by an audible alarm or directly to the unit to alert health care providers of a potentially harmful situation requiring immediate attention. These automated computerized systems increase the level of efficiency and safety for patient care by monitoring and alerting health care providers to potential safety concerns.

Machine learning can be categorized into two areas: supervised machine learning and unsupervised machine learning. In supervised machine learning, the algorithms used by the computer are developed through exposure to previous examples. This means that the software uses prior knowledge to program the algorithms that are then applied to the new data (Alpaydin, 2014). Supervised machine learning in AES requires the input of multiple previously scored items to “train” the program. In this situation, the program develops algorithms that learn the scoring behaviours of humans by analyzing the previously scored responses.

In unsupervised machine learning, there are no previous examples for the software to work with so the algorithms identify similarities and attempt to group this information to apply to new data (Alpaydin, 2014). The unsupervised algorithms are considered less accurate than the algorithms developed in supervised machine learning, however, because there is no need for a previously scored training set of data, it is less expensive to use an unsupervised system (Gierl, Latifi, Lai, Boulais, & De Champlain, 2014). Unsupervised machine learning systems would be more appropriate for use in lower-stakes examinations or in formative assessments whereas supervised AES systems are currently used for higher-stakes assessments.

Scoring. The AES systems considered for use in this study are supervised and require the previously scored items, as the training set, to build the models used for scoring the new data. Once the program builds the model, the model can then be applied to new data to analyze and score the responses. The AES system uses the information from the already scored items to build a model that provides the foundations for scoring future written responses. The program follows the algorithm models and applies that learning to score new data.

WEKA—An Automated Essay Scoring System

One of the challenges of using computer programs for education is the availability of the desired programs. WEKA, the acronym for Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand. It was created for the purpose of identifying information from raw data to complete several data mining tasks such as: classification, clustering regression, feature selection, visualization, and data preprocessing (WEKA, 2015). One of the capabilities of this computer program is automated essay scoring.

WEKA is an open source, supported computer program that can be used with almost every computer platform. It is in Java script which makes it easy to use and supported by most software platforms. WEKA is available for download on the WEKA Data Mining website. It is supported with current updates available in the package manager icon for the program.

To use WEKA for AES, the program is downloaded onto a windows 64-bit platform and the updated packages are installed. Selecting the package manager option ensures that the most up to date programming is installed. WEKA has four options for the users which include: 1) Simple—allows some features for the data; 2) Explorer—requires a graphical interface for experimentation on raw data; 3) Experimenter—allows users to conduct experimental variations on data sets; and 4) Knowledge Flow—which functions similar to the Explorer option but has some additional functionalities. For the purposes of this research project, the option of Experimenter is recommended (WEKA, 2015).

Once the Experimenter option is selected, there are six steps to process the raw data in order for them to be scored. These six steps are: 1) Preprocess—used to choose the data file for the application; 2) Classify—used to test and train different learning schemes on the preprocessed data file; 3) Cluster—used to identify clusters or groups within the data file; 4) Association—used to apply rules that identify associations within the data; 5) Select Attributes—applies different rules to see changes based on inclusion and exclusion criteria within the data; and 6) Visualize—used to identify what the manipulation outcome on the data is.

The data for preprocessing must be input as a CSV file. The responses from the students have been recorded on an excel spreadsheet which can be converted to a CSV file. The program recognizes an open file format to capture the data. Once the data is uploaded to the program, the experimenter can move through the remaining five steps.

The next step, classify, allows the experimenter to select a few options regarding the training set, cross validation, and split percentage options. Typically, the program will use 40 percent of the data to develop the training set. In a data set of 400 responses, the program will use 160 responses to develop the training set. Other options can be selected in this step as well, such as identifying specific attributes or rules for the program to follow. For example, if the students MUST include a specific term in their responses, this is the step that information can be stipulated for the program. The data set in this study does not require any rules to be selected. There are no set terms or words that the students must include in their responses. Nor are there any words or terms that the students cannot use in their responses. The program can use all available information from the data set without any classification rules.

The cluster step follows preprocessing and classifying. This is the process in the program that identifies commonalities within the data to produce information used for analysis. As in the classify step, the user can select the options for the training set in this step. The training set is 160 responses from which the program will identify commonalities, groupings, and clusters of occurrences in the data set. These commonalities or clusters will be used to identify the associations and yield the results.

Once the cluster step is complete, the associations step can be initiated to yield the results of the scoring. This is the step that the results of the scoring are determined.

The fifth step, select attributes, allows the user to select specific attributes of the data for calculation purposes. This is an additional step that can be used if the examiner wants to implement additional rules to the classifying of the data. If this step is not initiated, the program will default to using all of the available attributes identified in the evaluation of the data set.

The final step, visualization, is used to show the results in a selected pattern (such as scatter plot, grid, or graph) to give a visual representation of the results of the calculations. There are several options the user can select for these representations, but the most common are scatter plots or grids.

LightSIDE

Another AES system that is well known is the machine learning environment (MLE) called Light Summarization Integrated Development Environment (LightSIDE), Version 2.1.2. This is a supervised machine learning AES system. The process of using LightSIDE to score these items includes the following steps. First, the grouped responses to item #1 are exported to a CSV file which is uploaded to the LightSIDE program. The program runs the data through three modules: 1) data transformation—feature extraction module; 2) model building and evaluation module; and 3) automated scoring—absolute score prediction (Latifi, Gierl, & Boulais, 2013). The AES program requires 40 percent of the total responses to be used for “training the computer”. For example, if there are 400 written responses to score, the program will use 160 of the 400 responses to each item to “train the computer” and then automatically score the remaining 240. This means that the initial 160 responses in the training set must be previously scored. For the initial training responses, the computer will randomly select 160 responses with the corresponding scores from human rater #1 and the feature extraction module will begin. This is the process that transforms the examinee responses to establish statistical relationships between elements-of-text and human scores (Latifi, Gierl, & Boulais, 2013).

The processes of using WEKA and LightSIDE programs are similar. The main difference is the fact that WEKA is still an open source supported program which is updated on a frequent basis and includes community discussion groups and technical support. LightSIDE is

no longer open sourced or supported for “public” use. This is an important factor when selecting an AES program to use for current and future scoring.

Advantages of Automated Essay Scoring

Automated essay scoring has several advantages over human scoring. The most obvious is the savings in time and cost. Essay scoring by humans is such a costly and timely activity that it can result in examiners avoiding inclusion of essay-type items on examinations (Dikli, 2006; Williamson, Xi, & Breyer, 2012). In contrast, automated essay scoring systems can score thousands of essay-type items in seconds. Another advantage of AES is the objective scoring of items (Attali, Lewis, & Steier, 2012; Williamson, Xi, & Breyer, 2012). Human raters by nature are subjectively influenced when scoring items. Factors such as previous performance by examinees, fatigue, and general mood can influence the scoring by human raters. Also, human scoring of essays has been reported to have challenges in reliability. Several studies have shown that scoring by human raters can “drift” over time, meaning that how one human scored the first response to an item may not be the same as how that same human scored the 50th response to an item (Almond, 2014; Tan, Kim, Paek, & Siang, 2009). Also, reliability between two or more raters can be challenging and result in dramatically different scores for essays. The impact of inter-rater reliability is essentially non-existent in AES.

Another significant advantage of AES is the relative immediacy that students receive feedback about their performance (Dikli, 2006; Gierl et al 2014; Reilly, Stafford, Williams, & Corliss, 2014; Williamson, Xi, & Breyer, 2012). When written assessments are scored by human raters, there can be a long delay in the students receiving feedback about their performance. This delay can impact the effectiveness of their learning (Clynes & Raftery, 2008; Corrigan & Hadham, 2011; Cushing, Abbott, Lothian, Hall, & Westwood, 2011). Automated essay scoring

provides examinees with almost immediate feedback on their performance, which has been shown to increase the effectiveness of learning (Clynes & Raftery, 2008; Corrigan & Hadham, 2011; Cushing, Abbott, Lothian, Hall, & Westwood, 2011) and meet the increasing demands for rapid feedback in educational assessments (McNamara, Crossley, Roscoe, Allen, & Dai, 2015).

Human Raters and Automated Essay Scoring

Several studies have reported that scoring by AES and humans are highly correlated (Attali, Lewis, & Steier, 2012; Dikli, 2006; Gierl et al, 2014; Ramieni & Williamson, 2013; Walkow & Reilly, 2014), with some of these studies showing that correlations between AES and human ratings are as high as 0.97 (Attali, Lewis, & Steier, 2012; Attali & Burstein, 2006; Gierl et al, 2014). Automated essay scoring systems have been very successful in attaining levels of accuracy that are often as high as expert human raters (McNamara et al, 2015).

In previous studies, human raters are considered the gold standard for which the AES systems are compared with (Attali, Lewis, & Steier, 2012; Ramieni & Williamson, 2013; Shermis, 2014; Walkow & Reilly, 2014). Scoring is typically completed by humans so it is understandable that research studies use human raters as the standard for comparison when evaluating the effectiveness of AES. Ramieni and Williamson (2013) noted that scoring completed by humans is the most common criterion against which AES is evaluated.

Disadvantages of Automated Essay Scoring

There are also disadvantages to AES. Automated essay scoring systems require large numbers of practice responses to train the computer to score the item (Dikli, 2006). In situations where the total number of assessment to score is 60, the training set would be 40 percent of the total or 24 assessments. This means that a human would need to score the 24 essays and then

input these results into the program to score the remaining 36 essays. The accuracy of the algorithms developed from a small sample of essays (24) would not be as high as a larger number for the training set. In some cases, depending on complexity and variation of the written responses, the algorithms require much larger numbers in the training set to even run the program (Alpaydin, 2014). Therefore, AES may not even be possible to use if the total numbers of assessments are not large enough. Also, the examiner may decide to just go ahead and score the remaining 36 items rather than go through the processes of data preparation and input to AES programs.

Another disadvantage is AES requires computer hardware and software programs to score the items. Examinees must have access to computers to key in their responses and examiners must have access and ability to use the AES systems. This is important to note since the written responses need to be keyed into a computer—not handwritten. If the responses are handwritten, they would all have to be accurately transcribed to a file for input to the program. This obviously would be very time-consuming and negate the advantages of efficiency for this process.

Challenges to Automated Essay Scoring

There are a few challenges to AES and its utilization in scoring written assessments in all fields of study. One of these challenges results from the difficulty in accessing AES systems. Some of the AES programs are not available for public use and are limited by copyright or proprietary rights (Shermis & Morgan, 2016). This results in decreased awareness of AES, limited utilization of AES, and reduced research opportunities involving AES. All of these contributing factors result in decreased understanding, awareness, and therefore acceptance of AES as a potential strategy to score written assessments (Shermis & Morgan, 2016). As outlined

previously, it is human nature to resist the use of machines (including computers) to perform tasks that are typically done by humans. Pay parking machines, automated phone answering systems, auto fill, and writing recognition programs for cheque deposits through smartphones are all examples of change that were resisted until the effectiveness of the machines was proven. With the restrictions placed on the availability of AES to conduct research, this results in decreased awareness and limited opportunity to explore and gather information about AES.

Another significant challenge for AES is the perspective that machines cannot possibly mark assignments as well as humans (Shermis & Burstein, 2013) and that using AES systems lack human interaction (Dikli, 2006). Many people have the perspective that computers cannot understand the essence or context within a written assignment. Condon (2013) noted that the lack of human interaction and ability to assess the overall quality of the essay presents a significant challenge to the acceptance of AES. Although several studies have shown that AES systems score assessments very similar to human raters, there is still a lack of acceptance that AES is as good as human raters when, in fact, studies have shown that AES may actually be better than human raters (Shermis, 2014).

Finally, it is important to note that sometimes humans just resist change. This inertia is possibly the most challenging factor to acceptance of AES. Rather than look at options for scoring large numbers of written responses, constructed-response assessment items have been removed from courses. Essay scoring by humans is such a costly and timely activity that it can result in examiners avoiding inclusion of essay-type items on examinations (Dikli, 2006; Williamson, Xi, & Breyer, 2012). Written assessments have essentially been eliminated in courses with large class sizes and only implemented in courses with smaller numbers of students. This was evident in the search for data for this study. Written assessments for large class sizes

have been replaced with selected-response item assessments that are easier to score. This is a concerning trend since the literature supports the inclusion of written assessments to evaluate higher-level thinking skills in all disciplines, specifically nursing education.

Literature Review Conclusions

A literature search through databases such as ERIC, OVID, CINAHL, EBSCO, and MedLine found no examples of studies or articles related to the use or consideration of AES in nursing education. From these searches, it appears that AES has not been used for scoring any assessments in the field of nursing education.

As outlined in the literature review, effectively assessing higher-level thinking skills in nursing students improves clinical practice to ensure patient safety. Is AES an effective method to score constructed-response items to assess critical thinking, clinical reasoning, and clinical judgement in nursing students? Is AES as effective as human raters for scoring a constructed-response item in nursing education? The following research study examines the use of AES in assessments of critical thinking, clinical reasoning, and clinical judgement in nursing education.

Chapter 3: Methods and Data Collection

The following is an overview of the methods, data collection, and proposed analysis for this study. This chapter is divided into two sections. The first section includes a description of the processes used to identify and collect the data for the study. Descriptions of item development and administration processes, scoring procedures (including scoring rubric and human raters), ethics overview, and the automated essay scoring framework (WEKA) are included. The second section is an overview of the proposed analyses for the collected data including reliability coefficients and agreement measures.

Section One: Methods of Data Selection and Collection

A secondary analysis study is proposed to better understand the potential use of constructed-response items scored by AES for evaluation of critical thinking skills in pre-licensure nursing students utilizing a patient scenario. This secondary analysis design is a powerful measurement to identify significant differences in the scoring of responses by humans and AES (Gamst, Meyers, & Guarino, 2008).

Constructed- Response Items

As outlined in chapter one of this dissertation, multiple nursing educators across North America were contacted to obtain responses to constructed-response items for this study. A list of the programs that were contacted is included in Appendix A (see Appendix A). Automated essay scoring systems require large amount of responses to train the AES system to score the data. The researcher was unable to obtain enough responses from any of the contacts listed. Several of the programs use constructed-response items for smaller class sizes (about 20 students), but there were no examples of constructed-response items used for large class sizes. It

became apparent that, in order to explore the effectiveness of AES in nursing education, constructed-response items would have to be developed, administered, and scored.

The researcher met with three teaching teams at the University of Alberta, Faculty of Nursing (FON) to discuss and identify appropriate constructed-response items that would be relevant and accurate for learning assessments for nursing students working with patients in acute care settings. It was determined by the teaching teams that a patient scenario with several questions would be the best learning assessment for the students enrolled in the three courses. Candler (2007) outlined the value of using patient scenarios for education and learning assessments for higher-level thinking skills in nursing students. All members of the three teaching teams supported the use of a patient scenario and constructed-response items. The inclusion of constructed-response items for assessment was also supported by the undergraduate and FON leadership teams. There was verbal agreement that the FON needs to increase the use of constructed-response items in examinations and learning assessments.

One major issue that was discussed at the teaching team meetings was the scoring of the constructed-response items. Most of the teaching team members indicated that scoring these questions would take too much time for the number of students that would be completing the items.

Item Development. After meeting with the three teaching teams, the researcher began development of the patient scenario and items. An appropriate, realistic patient scenario was developed by the researcher and reviewed by all three teaching teams (see Appendix B). The teaching team members considered fairness, difficulty, and content relevance to ensure that the scenario was appropriate to include in a learning assessment for the students. Once all revisions to the scenario were completed (based on the feedback from the teaching teams), the

constructed-response items were developed. The constructed-response items were reviewed by three nursing experts (at the FON) in the fields of health assessment and nursing process before being reviewed by the teaching teams. Once all feedback had been discussed and incorporated into the scenario and constructed-response items, the items were given to volunteers to complete to ensure the accurateness of the questions. These volunteers are nursing colleagues in the FON. This process of item development is consistent with item development in learning assessments in the FON.

Once the items were developed and approved by the Associate Dean, Undergraduate Programs and the three teaching teams, the items were given to the Teaching and Learning Technologies (TLT) department at the FON. The teaching teams, researcher, and TLT worked together to develop a platform to administer constructed-response items on computer-based examinations. The platform that TLT developed is currently being used in multiple courses at the FON across several programs.

Upon completing the development of the scenario and items, a detailed scoring rubric was developed which included the weighting of each items. An identified concern with AES is the large amount of data needed to develop the corpus of text responses that the computer uses to score the constructed-response item. Often the literature supports having more data to build the corpora, however it is acknowledged that developing a substantive, comprehensive scoring rubric is also beneficial to the development of a corpus of text (Shermis & Burstein, 2013).

To ensure that an accurate, substantive, and comprehensive scoring rubric was developed, the constructed-response item (see Appendix B) was given to nine post-licensure nursing colleagues who volunteered to review and attempt to answer the item. The responses from these nine colleagues (see Appendix C) were compiled together. Their responses, along with the

researcher's responses were used to begin the development of the scoring rubric for the constructed-response item (see Appendix D). In addition, the researcher also reviewed the textbook and resources that the students used for learning the information which the questions were based upon and included this in the development of the scoring rubric. The responses to the constructed-response item submitted by the undergraduate nursing students were also included to develop the text corpus for the AES program. The teaching team leads reviewed the scoring rubric and weighting of each item based on the responses from the volunteer colleagues. This scoring rubric was followed to score the responses.

The patient scenario and constructed-response items were administered to approximately 400 nursing students at the FON and scored by a single human rater over the period of ten days. The responses and scores were collected and kept on file with TLT.

Ethics Approval

Application for ethics approval (see Appendix E) will be submitted online to the University of Alberta Research Ethics Office, specifically to the Research Ethics Board. This is the review board that administers the ethics review process for several faculties at the University of Alberta and has representation from the following faculties: Psychology, Educational Psychology, Business, Medicine, Agriculture (Food and Nutrition Science), and Physical Education and Recreation. The Research Ethics Board is governed by the University Committee on Human Research Ethics (UCHRE).

The application for approval will be submitted specifically to the Research Ethics Board committee number 2 (REB 2). Research Ethics Board 2 "reviews all interventional research designs including (but not limited to) training interventions for educational, psychological, social or performance purposes" (Research Ethics Office, 2016). Research Ethics Board 2 also

“reviews all research where the primary ethic concern is privacy and/or confidentiality” and includes access to University of Alberta students (Research Ethics Office, 2016).

This research study requires access to constructed-responses submitted by University of Alberta students in the Faculty of Nursing. Access to the actual students is not required but access to the responses that students put on their examinations is required. Therefore, REB 2 is the appropriate ethics board from which to request approval.

Selection of Data

Following ethics approval from the University of Alberta Research Ethics Committee, responses to short-answer items previously administered on nursing education examinations involving a patient scenario will be collected and analyzed (see Appendix B). The student responses to the questions have already been used in examinations within the University of Alberta Bachelor of Science in Nursing Program and ethics approval is requested to access and analyze the already recorded responses. Literature in AES outlines the importance of collecting and analyzing substantial numbers of responses in order to “train” the computer to score the responses (Shermis & Burstein, 2013). The goal for the number of responses collected and analyzed in this study to “train” the computer is 400. Multiple attempts to access larger numbers of data were unsuccessful (see Appendix A). Several nursing programs and other health sciences programs across North America were contacted and none of these programs had sufficient numbers of responses to short-answer examination items. However, essentially every faculty member from across North America that was contacted in regards to this study expressed high levels of interest in AES and the potential for its use in nursing education and other programs. A common comment from the faculty members contacted was a request to “please let them know if this works so they can put short answer questions back onto their examinations for students”.

The number of responses that were accessible for this study was limited to 400 which is considered a substantive amount but still less than the “vast number” of responses that would be optimal (Shermis & Burstein, 2013). A detailed scoring rubric for the constructed-response item was developed (see Appendix D).

The questions that were selected for scoring by AES are related to a patient scenario that is commonly seen in nursing practice and patient care areas. The chosen scenario requires the students to accurately collect information from the patient using health history questions, physical examination techniques, diagnostic tests, and then identify possible interventions to help the patient (see Appendix B). This patient scenario and questions require students to critically think and use clinical reasoning and judgement to outline critical information collection and safe aspects of patient care for the patient in the scenario.

Data Collection Procedures

Short-essay responses to items involving a patient scenario will be collected from the undergraduate examination storage files at the University of Alberta, Faculty of Nursing (see Appendix C). Permission to access the stored responses will be requested from the Vice Dean, Faculty of Nursing at the University of Alberta (see Appendix F). The 400 responses to each item have been scored by human rater #1 and these scores have been recorded and stored with the Faculty of Nursing. The responses will be assigned a case number from 001-400 in random order. No identifying data of the students will be collected. None of the actual responses will have any identifying data and all responses will be pooled into one complete data file to ensure complete anonymity of the responses.

In order to determine the effectiveness of AES, the scores generated by AES will be compared with the scores assigned by two human raters. These human raters will independently

score 400 responses and record the scores for each response. The human raters selected for this process will have expertise in the content area (as determined through prior research and published contributions), but will not have been directly involved in teaching the students who have completed the examination. This will further ensure anonymity and protection of the students' responses. Following the scoring by the human raters, the responses will be scored by AES and the scores will be compared.

In order to export the data to a usable file for AES, the responses will be grouped by item. That is, all 400 responses to item #1 will be pooled together in a data file for this study. Then another data file with all 400 responses for item #2 will be pooled together, and so on for items #3 and #4. Each response will be assigned an identifying number of 001-400. These numbers will be randomly assigned to the responses and will not be related to student names or any identifying data. Any identifying data will not be exported with the responses. Only the responses and scores will be exported. The process of "training the computer" requires the data file to have the scores associated with the responses. The responses from the archived test responses have previously been scored and these scores are recorded with each response. The data for preprocessing must be input as a CSV file. The responses from the students have been recorded on an excel spreadsheet which can be converted to a CSV file.

The AES system chosen for this project is the machine learning environment (MLE) called WEKA, Waikato Environment for Knowledge Learning developed at the University of Waikato in New Zealand (Witten, Frank, & Hall, 2011). The processes for WEKA are previously outlined in chapter two of this dissertation. WEKA was selected as the AES system for this project because it is an open-sourced supported program that is updated on a frequent basis and includes community discussion groups and technical support. Although LightSIDE is

also an appropriate AES system for this study, it is no longer open-sourced or supported for public use.

The WEKA program recognizes an open file format to capture the data which can then be run through the program. The AES program requires 40 percent of the total responses to be used for “training the computer”. For this research project, the program will use 160 of the 400 responses to each item to “train the computer” and then automatically score the remaining 240. For the initial training responses, the computer will randomly select 160 responses with the corresponding scores from human rater #1 and the feature extraction module will begin. This is the process that transforms the examinee responses to establish statistical relationships between elements-of-text and human scores (Latifi, Gierl, & Boulais, 2013).

Once the data is uploaded to the AES system, the uploaded data will go through the AES processes of data preparation and preprocessing, feature extraction, model building, and scoring. As outlined in chapter two, the specific steps for the WEKA program include: preprocess, classify, cluster, associations, select attributes, and visualization to score the responses.

Data preparation and preprocessing are critical first steps for the AES system to run. The steps previously outlined for uploading the data allow the program to identify and transform the text elements into units that the program can work with. Once the preprocessing step is complete, the program will use algorithms to identify features and extract these commonalities to essentially reduce the amount of data to a usable amount. In essence, the program identifies the common important features within the text responses to decrease the volume of data needing analysis. Then the program will use these identified features to build a scoring model to apply to new data and then score the new data. These steps are outlined in more detail in chapter two.

For the purposes of this research study, a second human rater will be employed to independently score the responses using the scoring rubric developed (see Appendix D). The scores of human rater #1 have already been recorded and input into WEKA. The scores of human rater #2 will be recorded and scored in a separate data file. The data file for human rater #2 will be converted to a CSV file and uploaded to the AES and run as a set of new responses. The scores from human rater #2 will not go through the same processes as the data file from human rater #1. The 240 responses from human rater #1 and 400 responses from human rater #2 will be scored as new responses by the AES program. The AES program will record the scores of the new responses and then the scores from human rater #1, human rater #2, and AES will be analyzed and compared.

Section Two: Analysis Procedures

This section includes an overview of the proposed analysis procedures for this research study. Reliability coefficients and agreement measures are proposed as relevant analysis procedures.

Automated essay scoring systems are generally designed to reflect human ratings (Attali & Burstein, 2006; Dikli, 2006; Shermis & Burstein, 2003). With this consideration in mind, the analysis will focus on comparing the scores generated by human raters with AES. Results from human rater #1, human rater #2, and AES will be analyzed to determine significant differences in scoring. The analysis will include the following comparisons: human rater #1 with human rater #2; human rater #1 with AES; and human rater #2 with AES.

Reliability Coefficients

Score correlations will be analyzed using reliability coefficients calculations for the following three comparisons:

- 1) Human rater # 1 compared with human rater # 2;
- 2) human rater #1 compared with AES; and
- 3) human rater #2 compared with AES.

Reliability refers to the consistency of test scores (Shermis & Burstein, 2014). Specific reliability information for this study would be the consistency between human raters and each human rater with AES. In this study, reliability coefficients will be calculated for all three comparisons. Some AES studies report as high as 0.97 coefficient results comparing human raters to AES (Attali & Burstein, 2006). Conversely, human to human score coefficients have been reported to be much lower (Shermis & Burstein, 2014). Reliability coefficients results at 1.0 indicate perfect agreement whereas, 0.0 indicate no agreement. Ideally, coefficient results over .80 indicate a high level of consistency for the scores (Shermis & Burstein, 2014).

Williamson, Xi, & Breyer (2012) noted that correlation coefficients comparing human and machine scoring must be over .70. The specific validity coefficients that will be calculated and analyzed are as follows: inter-rater reliability coefficients between human raters and machine scoring. Pearson correlation coefficients comparing human rater #1 with human rater #2; human rater #1 with AES; and human rater #2 with AES will be calculated analyzed to determine how closely AES and human raters score the items.

Agreement Measures

Several agreement measures will be used to analyze the results. Exact-agreement percentage, adjacent-agreement percentage, Cohen's Kappa κ , and Cohen's Quadratic Weighted Kappa κ_q will be calculated and analyzed.

Exact-agreement measures and adjacent-agreement percentage measures refer to the agreement between two scores. Exact-agreement measures look at the agreement between two

scores as being the same. This means that the two raters give the same score to the response. For example, human rater #1 and human rater #2 both give the score of 82 percent to the item being scored. The adjacent-agreement percentage measures refer to the agreement between scores as being within one point of each other (Shermis, 2014). This means that the two raters give two scores that are within one point of the other. For example, human rater #1 gives the score of 82 percent and human rater #2 gives the score of 83 percent. An important consideration is that the unit of measurement for these calculations will be the unit input to the program. If the score is recorded as 18/20, then 17/20 or 19/20 will be considered acceptable in terms of adjacent-agreement measures.

Both exact-agreement and adjacent-agreement measures are calculated as a percentage and reported as κ . A κ of 1.0 indicates perfect agreement whereas a κ of 0.0 indicates no agreement. Cohen's Kappa measures the agreement percentages between two raters (human to human or human to machine). Cohen's Kappa is identified as a stringent measure because it corrects for the likelihood that some agreement between raters occurs by chance (Graham, Milanowski, & Miller, 2012).

It is important to include inter-rater agreement analysis in this study to investigate the effectiveness of using AES for assessment of student performance. Human raters can have high-reliability measures yet be low on agreement measures. By including all these analyses, more information on the effectiveness of AES in nursing education assessments can be determined.

Conclusion

The advancement of technology has dramatically changed the way we teach and learn and has given us different learning strategies to incorporate into education programs. The Canadian Patient Safety Institute (2008) recognized that with all the technological advancements

available to health care professionals, it is no longer acceptable to “practice on patients”. Patient scenarios are one method of teaching and assessing the development of critical thinking, clinical reasoning, and clinical judgement skills in nursing students. The use of constructed-response item format questions to assess higher cognitive functioning can result in many challenges for educators. Automated essay scoring may be one solution to the challenges associated with scoring multiple constructed-response items. Critical thinking, clinical reasoning, and clinical judgement are essential skills to develop in health care professionals and are viewed as critical to patient safety. The proposed research study will provide information on the effectiveness of AES to score assessments on critical thinking, clinical reasoning, and clinical decision-making skills in nursing education. Automated essay scoring may provide a strategy to allow the increased use of constructed-response items which generates feedback for students to develop their critical thinking skills which, ultimately improves clinical practice and patient safety.

Chapter 4: Results

This chapter is organized into five sections. The first section is an overview of the processes used to obtain the data including the additional processes implemented to ensure ethical and accurate implementation of the study. It also includes examples of the responses from the four items scored by human rater #1 (HR1) and human rater #2 (HR2) and an overview of the data export processes. The second section outlines the implementation of the automated essay scoring (AES) software program and the results from AES of the four items for the cases in this study. The third section of this chapter presents the findings from the comparisons between HR1, HR2, and AES. The fourth section outlines the processes and results for data collected in June 2018 and application of the AES model developed with the initial data using the same comparisons as section three. The fifth section in this chapter includes a summary of the results for all comparisons and items.

Section One: Overview of Data Collection Processes and Export of Data Files

Following ethics approval received from the Research Ethics Board at the University of Alberta Research Ethics Office (see Appendix E), permission to access the data was requested and granted by Dr. Joanne Profetto-McGrath, Vice Dean, Faculty of Nursing, University of Alberta (see Appendix F). Once ethics approval and permission to access the data were granted, the processes for data export were initiated.

The items and responses used in this research study were from computer-based examinations in three different courses at the Faculty of Nursing. The total number of student responses (across the three courses) was 359. All four items were administered to 359 students through computer-based examinations on *eclass*, the course-based delivery platform used at the University of Alberta. The 359 responses for the four items were recorded and exported to Excel

spreadsheets. No student identifying data were exported and each case was assigned a case number. The following images are examples of the data files for the responses scored by HR1 for item #1 (see Figure 4.1), item #2 (see Figure 4.2), item #3 (see Figure 4.3), and item #4 (see Figure 4.4).

Case	Response to Item 1	HR #1
1	1.Does she have hypertension? 2.Does she have a family history of stroke? 3.Does she smoke? 4. What is her admitting diagnosis to the hospital? 5. What is her BMI (is she overweight or obese?) 6. Has she had a stroke/CVA before? Or any complications of embolism formation? 7. Is she physically active? 8. What medications is she on (birth control? anticoagulants? antihypertensives?) 9. Her baseline vital signs 10. What is her ethnicity/background? 11. What other co-morbidities does she have? (dyslipidemia, diabetes, Coronary Artery Disease)?	HR #1
2	12. What does her diet typically look like? (Is it high in processed foods and fats?) * When did it start? * What makes it feel better? * What did you have to help? * Onset of the pain? * location of the pain? * Duration of the pain? * Character of the pain? ("explode") * Any medications that were taken recently? * Rate the pain? (0-10) * Severity of the pain? * Did you fall? What were you doing? * What was the patients admitting diagnosis? * What caused you to come to the hospital? (what made her walk into the hospital?) * The patient have any allergies?	2

Figure 4.1. Example of responses for item #1 scored by HR1.

Case	Response for Item 2	HR #1								
1	<p>1. Respiratory Rate</p> <p>2. Pulse</p> <p>3. Oxygen Saturation</p> <p>4. Blood Pressure</p> <p>5. Temperature</p> <p>6. Facial expression (is there a droop?)</p> <p>7. Does she have any slurred speech?</p> <p>8. Can she lift both arms equally above her head</p>									
2	<p>1. Look at pupil response (PEARL?)</p> <p>I would preform:</p> <ul style="list-style-type: none"> * Vital signs * Check for alleviating pain * Check the skin colour * Check the skin temperature * Check for any lumps or bumps in the head * Check for any bleeding or cuts in the head * Palpate the head and check for symmetry as well as if any patches of hair missing * Check for dehydration. This could lead to patient having the headache. Check moisture of the mouth and skin, the colour of the urine. * Check for anything else which could lead the patient having a headache. Such as hypoglycaemic if the patient is diabetic and etc. 	1.25								
3	<p>2 * Check for any bruises which could result from a fall</p>	1.5								

Figure 4.2. Example of responses for item #2 scored by HR1.

Case	Response for item 3	HR #1
1	1. CT Scan 2. MRI 3. X-ray 4. Blood Test (especially platelet and RBC count will be lowered if there is bleeding in the case of a haemorrhagic stroke)	
2	1 - look at electrolytes as well (are there any abnormalities?) * MRI	1
3	2 * Blood test * CT scan * MRI	1
4	3 * Blood work CBC 1. CT scan to assess the condition of the brain, possible stroke 2. MRI to further assess the brain	1
5	4 3. Culture and sensitivity **CT, MRI, angiogram to determine if there is a stroke and if it is hemorrhagic or occlusive ultrasound of brain bloodwork to assess platelet level, electrolyte levels, albumin, WBC count, hematocrit and hemoglobin levels for bleeding interarterial blood pressure for a more accurate BP reading	1
6	5 urinalysis to identify abnormalities	1

Figure 4.3. Example of responses for item #3 scored by HR1.

	A	B	C	D	E	F	G	H	I
1	Case	Response to Item 4	HR #1						
		If it is an ischemic stroke: You want to keep the blood pressure as low and regular as possible 1. Administer TPA right away if still in the window of time (1 hour would be) -- this would more be notifying the physician and then the nurse would administer it 2. Lay flat and keep in bed (perform any required ADLs required for them) 3. Give analgesic if in pain (blood pressure/breathing/pulse may increase in event of severe pain) 4. Reduce environmental stimulation (dark room, quiet, no visitors, no roomates, no TV/Music) 5. Administer laxatives (reduces straining- need to keep blood pressure as low and stable as possible) 6. Educate patient on situation/diagnosis to reduce anxiety/fears 7. (Once no longer in danger)- Encourage patient to perform exercises for rehabilitation as suggested by the physical therapist 8. Give anticoagulants (warfarin) to prevent another clot (this is not the case in haemorrhagic strokes) 9. Refer to occupational therapist before discharge- certain ADLs may be more difficult to perform independently 10. Encourage primary prevention of another stroke: educate the client on the importance of smoking cessation and proper nutrition and exercise, as well as the importance of taking prescribed medications on time							
2	1	* Administer medications to help relieve the pain (acetaminophen) * Get her to rest in bed (sleep) * Reduce the surrounding noises which may help the pain get worse * Have the patient get her mind of this by distracting her with other things to do * Reduce risk of falls; if the patient gets dizzy and falls and her hurts herself more * Get the patient to go for a walk * Have the patient have family visit. She may feel lonely. * Increase the patients water(fluid)intake. So the patient isn't dehydrated. * Increase physical activity so the patient stays active and doesn't have headaches	3.5						
3	2	* Manage other co-morbidities so the patient doesn't result in having side effects such as a headache. * Administer non-opioid analgesics * Dim the lighting * Remove loud noises from room * Distraction from pain	2						

Figure 4.4. Example of responses for item #4 scored by HR1.

Computer-based examinations administered through the *eclass* platform are compatible to export to Excel spreadsheets for data analysis. The total amount of time that HR1 required to score all 4 items for 359 students was 34 hours. This averages 5.75 minutes to score each student’s responses to the 4 items.

Scoring the Responses by the Second Human Rater

Once the responses were exported into data files, the responses then needed to be scored by a second human rater. In order for the second human rater to score the items that were

already scored and also to ensure the same processes were followed, it was necessary to create a “dummy class” on the *eclass* platform. This ensured that HR2 followed the same processes for scoring as HR1 in relation to reading the responses in exactly the same format, font, size, information, location, and aspects of computer-based examination scoring (such as automatic calculation of total scores and word counts). Also, by creating the “dummy class”, HR2 had no access or possibility of viewing the scores given by HR1 ensuring the scores were independent.

The second human rater followed the scoring rubric developed for the four items (see Appendix D) as a guideline and completed the scoring. The scores by HR2 were recorded within the *eclass* platform and then exported to an Excel spreadsheet. The total amount of time required by HR2 to score all 4 items for 359 students was 39 hours. This averages to 6.50 minutes to score the 4 items for each student’s responses.

The scores from HR2 were aligned with the student responses. To ensure that the case numbers and responses were accurately matched, the researcher along with an assistant, checked the responses and case numbers scored by the first human rater with the responses and case numbers scored by the second human rater to ensure the case numbers accurately matched the responses. It was concluded that the case numbers and responses were perfectly matched in the data files. For future studies with these types of files, it would be helpful to assign case numbers to the responses before exporting the responses to eliminate these extra steps. For this study, the responses and scores were exported before a case number was assigned which resulted in requiring extra checks to ensure that the case numbers accurately corresponded to the responses.

Once it was determined that the case numbers and responses were accurately aligned, the scores assigned by HR2 were exported into an Excel spreadsheet with the corresponding case numbers. Since the responses scored by HR2 will not be scored by AES, only the case number

and scores were exported to the data file for comparisons. This is the rationale for ensuring that the responses and case numbers were accurately matched. The following images are examples of the data files for the scores by HR1 and HR2 for item #1 (see Figure 4.5), item #2 (see Figure 4.6), item #3 (see Figure 4.7) and item #4 (see Figure 4.8).

Case	HR #1	Human Rater #2	AES
1	2	2.5	
2	3	3	
3	2.75	3	
4	1.5	3	
5	2	3	
6	2.5	2.5	
7	1.5	3	
8	3	3	
9	2.5	3	
10	2	3	
11	3	3	
12	0.5	1.5	
13	3	3	
14	1.5	2.75	
15	2.5	3	
16	3	3	
17	3	3	
18	1.75	3	
19	3	3	
20	3	3	
21	3	3	
22	3	3	
23	3	3	
24	1	2.75	
25	3	3	
26	3	3	
27	1.75	3	
28	2.5	2.75	
29	3	3	
30	3	3	
31	2.5	3	
32	2.5	3	

Figure 4.5. Example of scores for item #1 by case number for HR1 and HR2.

Case	HR #1	Human Rater #2	AES
1	1.25	1.75	
2	1.5	2	
3	1.25	1.25	
4	2	2	
5	2	2	
6	2	2	
7	2	2	
8	2	2	
9	2	2	
10	2	2	
11	0.75	1.25	
12	2	2	
13	1	1.75	
14	1.75	2	
15	1	2	
16	1.75	2	
17	0	2	
18	2	2	
19	2	2	
20	1.25	1.75	
21	2	2	
22	1.75	2	
23	1.75	2	
24	1.75	1.75	
25	1.5	2	
26	1.25	1.75	
27	2	2	
28	2	2	
29	2	2	
30	0.75	1.5	
31	1.75	2	
32	2	2	
33	1.25	1.75	

Figure 4.6. Example of scores for item #2 by case number for HR1 and HR2.

Case	HR #1	Human Rater #2	AES
1	1	1	
2	1	1	
3	1	1	
4	1	1	
5	1	1	
6	1	1	
7	1	1	
8	1	1	
9	1	1	
10	1	1	
11	1	1	
12	1	1	
13	1	1	
14	1	1	
15	1	1	
16	1	1	
17	1	1	
18	1	1	
19	1	1	
20	1	1	
21	1	1	
22	1	1	
23	1	1	
24	1	1	
25	1	1	
26	1	1	
27	1	1	
28	0.5	0.5	
29	1	1	
30	1	1	
31	1	1	
32	1	1	
33	1	1	

Figure 4.7. Example of scores for item #3 by case number for HR1 and HR2.

Case	HR #1	Human Rater #2	AES
1	3.5	4	
2	2	4	
3	2.5	3	
4	2.5	3.5	
5	3.5	3.5	
6	2	4	
7	2.5	4	
8	3.5	4	
9	2	3.5	
10	2	4	
11	2.5	4	
12	3	4	
13	2	2.5	
14	3	4	
15	4	4	
16	3	4	
17	4	4	
18	2	4	
19	2.5	4	
20	2.5	4	
21	3	4	
22	3.5	4	
23	2.5	4	
24	2.5	4	
25	2.5	4	
26	3.5	4	
27	3	4	
28	3.5	4	
29	4	4	
30	3	4	
31	1.5	3	
32	2	4	

Figure 4.8. Example of scores for item #4 by case number for HR1 and HR2.

Once the data files from both human raters were exported, the process of running AES with the responses from the students and the scores from HR1 to train the program was initiated.

Section Two: Results from Automated Essay Scoring Program

The Excel data files for HR1 required conversion for input to WEKA. The AES software requires data files to be in .csv or .arff file formats. Once the Excel spreadsheets were completed, the conversion and upload processes were initiated resulting in several challenges.

Conversion and Upload Processes

The Excel spreadsheets data for the responses scored by HR1 were converted to a .csv format then uploaded to WEKA. Initially, the files were uploaded to WEKA explorer before moving the files to experimenter. The first several attempts to upload the files to WEKA resulted in

multiple error messages due to the formatting of the files. In order to identify and correct these errors, the Excel spreadsheets were opened with sublime text to analyze the errors.

The first challenge with the formatting of the responses was the line breaks that students had inserted within their answers. For example, several students responded to item #1 by listing 12 different statements and putting a line break (enter) in between each text entry. The students keyed the enter (return) key after each statement which resulted in an error message for the upload. The WEKA program read this as 12 different entries rather than one entry. To overcome this challenge, the line breaks were removed by replacing all line breaks with a simple space. This did not alter any of the content or responses by the students, it just simply removed the line breaks within their answers and replaced these with a space. The example below shows how the response looked originally then after the line breaks were removed (see Figure 4.9). As outlined, none of the content was changed in the student's response in this example. The line breaks were replaced with a space and all content remains the same.

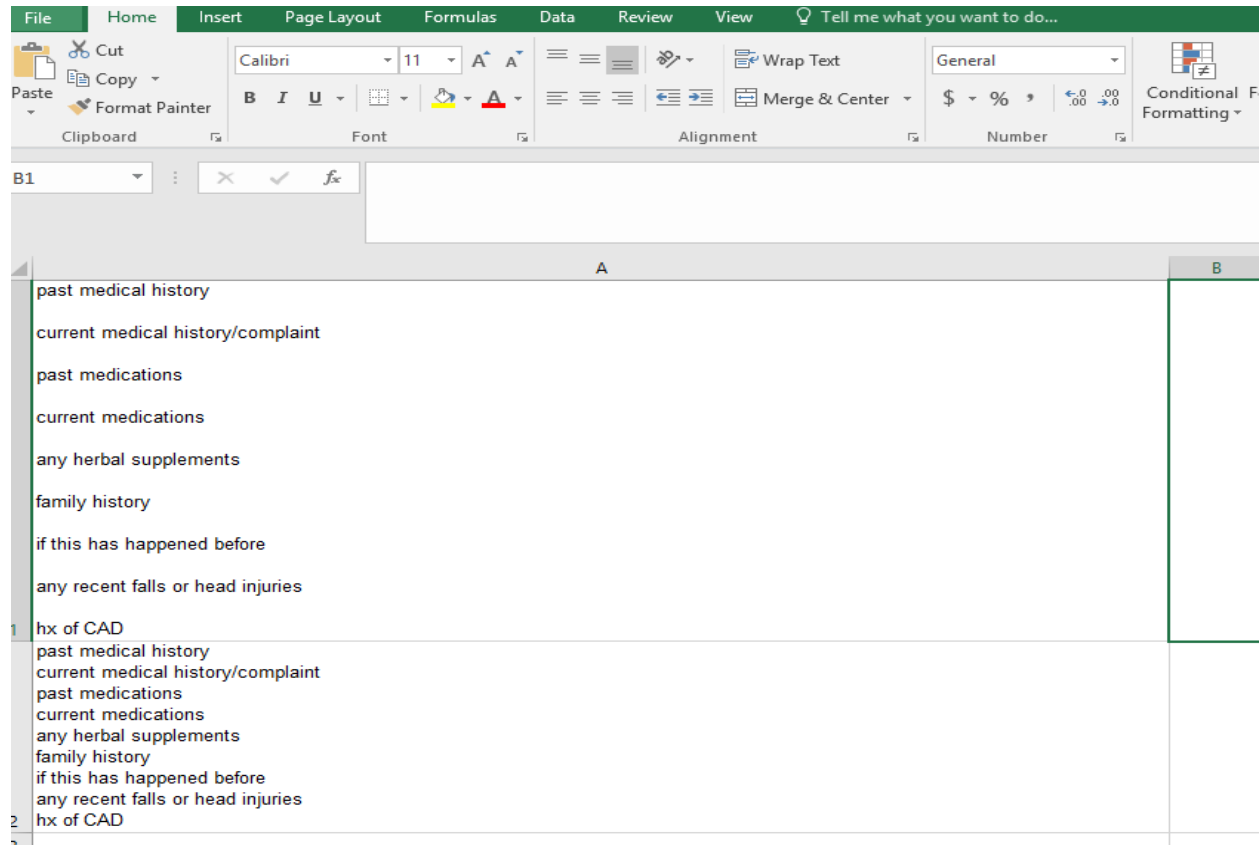


Figure 4.9. Example of line breaks removed from response and replaced with a space.

The second error preventing uploading the information was the use of dash space. Several students began their answers with a dash followed by a space then followed by the words (- Does the patient...). The software program recognized the dash followed by a space as a formula rather than the start of a new string of words. These needed to be adjusted to allow the upload of the file. Replacing dash space with a space corrected the issue and did not change any of the content in the responses. Several students began their responses with a dash followed immediately by a letter or word (-Does the patient...). These cases were accepted by the program. Only the instances when the dash was followed by a space required adjustment.

Another challenge was the coding that occurred when the Excel spreadsheet was converted to .csv format. In the Excel spreadsheet, there were several unused cells over multiple

columns. When the files were converted, each of the empty cells was represented with a comma which resulted in the presence of multiple commas following the content. This was easily adjusted by removing the additional empty columns from the Excel spreadsheets before converting to .csv format.

An additional challenge to uploading the files to WEKA was the occurrence of duplicate quotation marks. In some of the cases, students had inserted double or triple quotation marks. This was likely an error when students were keying their responses into the examination. Some of the students had additional key strokes that appeared as multiple quotation marks. The WEKA program identified these as errors and prevented the upload. By replacing all double keyed quotation marks with a single quotation mark, the error was corrected and none of the content was changed in the students' responses.

There were several other challenges to upload the files to WEKA which resulted in needing to clean the responses from all the key strokes that prevented the upload. To do this, the syntax called, "regular expressions", was used. This is a generic concept syntax across programming languages. The codes listed below outline what were cleaned from the data:

Codes:

remove newlines:

`~[\r\n]+~` maps to a space

remove hyphens:

`[-\s\s\^-|^-]` maps to a space

remove punctuation:

`[*\|?|!|\(|\)|:|\.|"|\'|=|,]` maps to space

[/] maps to space

remove numbered lists (1., 2., etc or 1), 2), etc)

[\d.\d{2}.] maps to space

[\d\)\d{2}\)] maps to space

To generate the results:

- 1) a filter runs on both the training set and test set to split each answer into words
- 2) a classifier (J48) runs on the training set to try and build an appropriate tree
- 3) the classifier (tree) is applied to the test set, which generates a score

Once all of these errors were adjusted, the file upload was successful for item #1.

However, the uploaded file of responses for item #1 was not recognized by the program and therefore could not be scored. At this point, the decision to try the data preprocessing program called Python was made. Python also has AES capability.

Data Preprocessing and Python

Data preprocessing is a critical step in AES and most data mining processes. It is an often overlooked and underappreciated essential step in preparing data for usage (Albon, 2018; Alpaydin, 2014; Witten, Frank, & Hall, 2013). Preprocessing the text data is essential for the software program to build the features of the model which is then applied to scoring data. It is especially important when using small data sizes (Bishop, 2006) which is the case for this study. Almost all text data needs to be cleaned and preprocessed before it can be used to build the features of a scoring model (Albon, 2018). Python uses standard string operations which can clean and preprocess most text. When several challenges prevented the use of WEKA for this study, Python was chosen as a possible solution for cleaning the text data and preprocessing it for upload.

Data preprocessing. Data preprocessing is a technique in data mining that essentially transforms the raw data into understandable formats for software programs (Albon, 2018; Techopedia, 2018). Raw data is often not understandable by software programs due to being incomplete, noisy, inconsistent, and containing errors and/or coding trends. Data preprocessing is proven to resolve these issues to transform raw data into understandable formats (Albon, 2018; Techopedia, 2018; Witten, Frank, & Hall, 2013).

As noted previously in this chapter, AES software must recognize the data for any upload and scoring. Even simple keystrokes that examinees include in their responses can prevent the upload or processing of the text. In order to prepare the data for AES to understand, recognize, and score, the data was preprocessed using the software program Python. Python uses operations which identify the areas of the text data requiring preprocessing which includes: data cleaning, data transformation, data reduction, and data discretization (Albon, 2018).

Data cleaning is the initial step in preprocessing which includes filling in missing values, smoothing out noisy data (data that has errors or outliers), and resolving inconsistent data such as inclusion of symbols instead of text. Once the data has been cleaned, the process of data transformation begins. Data transformation includes normalization of data to identify specific attributes and generalize the identified attributes to incorporate into the model. Data reduction is the next step in the preprocessing method that involves reducing the number of identified attributes to make the data more manageable to use. Data discretization involves reducing the values within the attributes so there is less data to process (Albon, 2018; Alpaydin, 2014; Techopedia, 2018; Witten, Frank, & Hall, 2013).

Once Python was downloaded and enabled, the process of cleaning, uploading, and preprocessing was initiated. The code used for the data preprocessing within Python for this

project was identified and created by Shin (2018) and is outlined in Appendix G.

Following the input of the developed code for data preprocessing, Python was able to preprocess the data and identify results. In the setup of the code, the commands for keywords were included for the scoring of the responses. This is quite similar to creating a scoring rubric with keywords and content outlined for human raters to follow. The sample size of 359 is considered small for AES (Shermis, 2013) which indicates the need for identifying the effective keywords from the scoring rubric and training set (Shermis & Burstein, 2013) and data preprocessing.

The keywords for item #1 were: ['aggravating', 'allergies', 'alleviate', 'associated', 'conditions', 'duration', 'family', 'headache', 'location', 'medications', 'OLDCARTS', 'pain', 'perspective', 'PQRST', 'PQRSTUV', 'quality', 'scale', 'significance', 'sleep', 'smoking', 'stroke', 'surgeries', 'timing', 'travel'].

The keywords for item #2 were: ['Cincinnati', 'drifting', 'facial', 'FAST', 'frown', 'gait', 'Glasgow', 'handgrip', 'LOC', 'neuro', 'neurovitals', 'paralysis', 'smile', 'speech', 'stroke', 'swallow', 'symmetry', 'visual'].

The keywords for item #3 were: ['bloodwork', 'MRI', 'CT', 'scan', 'urinalysis', 'ECG', 'CBC'].

The keywords for item #4 were: ['positioning', 'semifowlers', 'NPO', 'IV', 'intravenous', 'oxygen', 'O2', 'analgesics', 'siderails', 'safety', 'family', 'reassurance', 'support', 'TPA', 'aspirin', 'anticoagulant', 'calm'].

Development of the AES Model

In order to build an optimal scoring model for these four items, a training set of 80 percent was used and multiple runs were completed to identify the optimal scoring model. The

scale increments were very small which promotes a more stringent development of the scoring model which then can be applied to categories of scores (Bishop, 2006; Rudner & Liang, 2002). For example, item #1 was scored between 0 and 3.00 in 0.25 increments which are considered small increments (Bishop, 2006). By using these increments for the training set, the scoring model is developed with precise scaling. Then the scoring model can be applied to categories of scores. In a study by Rudner and Liang (2002), AES was implemented to successfully score three categories of responses (Appropriate, Partial, and Inappropriate) which supports the use of categories for AES.

Once the scoring model was developed from the training set, Python scored the remaining 20 percent. The following are the results for the comparisons from the AES model.

Section Three: Results from Comparisons between Human Raters and AES

In this section, the results from the comparisons between HR1, HR2, and AES are included. The data from the scores by HR1, HR2, and AES were input to SPSS v. 25 for statistical calculations for reliability coefficients, agreement measures, and discrepancy analysis.

Specifically the analysis of the comparisons were between:

- 1) HR1 compared with HR2;
- 2) HR1 compared with AES; and
- 3) HR2 compared with AES.

Agreement Measures and Reliability Coefficients

The agreement measures for the analysis of the data for the four items include exact percentage agreement and adjacent percentage agreement. Exact percentage agreement is the number of times the two raters scored the item exactly the same and is expressed as a percentage

value. A value of 1.00 means that the scores were all exactly the same whereas a value of 0 means that none of the scores were the same. These values are important to understand since the percentage of agreement indicates the similarity of the scores by the raters.

Adjacent percentage agreement is calculated by including the exact percentage agreement and also the scores that were within a determined score value. For example, item #1 was scored out of a possible total of 3.00 marks and was scored in increments of 0.25. To calculate the adjacent percentage agreement, the exact percentage scores and the scores that were within ± 0.25 for the raters were combined and expressed as a percentage value. A value of 1.00 means that the scores were either exactly the same or within 0.25 marks between raters. Agreement measures are different from correlations because they take into account the amount of agreement between raters. Correlations indicate associations but not necessarily agreement. The closer the values for agreement measures are to 1.00, the higher the agreement between raters. When determining what level of agreement is acceptable, Graham, Milanowski, and Miller (2012), Hartmann (1977), and Stemler (2004a), noted that 90 percent agreement is high and 75 percent agreement is the minimum standard that should be accepted.

The reliability coefficients calculated for the data include Cohen's Kappa κ , Quadratic Weighted Kappa κ_q , and Pearson r . Cohen's Kappa κ is considered a stringent measure for reliability between raters and takes into account the disagreement between two raters however, Cohen's Kappa κ does not take into account the degree of disagreement between the two raters (Graham, Milanowski, & Miller, 2012). In essence, Cohen's Kappa κ identifies whether there is disagreement between raters but not the amount of disagreement. To account for this issue, Cohen's Quadratic Weighted Kappa κ_q is used to identify the degree of disagreement between two raters. The quadratic weighted kappa κ_q is calculated using weights assigned to

disagreement. The higher the disagreement, the higher the weight which is calculated using a matrix computation. Generally, reliability coefficients between 0 and 0.20 indicate none to slight levels of agreement, 0.21 to 0.40 indicate weak agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate adequate to substantial agreement, and 0.81 to 1.00 indicate almost perfect agreement (Graham, Milanowski, & Miller, 2012; Field, 2018; McHugh, 2012). The guidelines for evaluating the κ_q and Pearson r values when comparing human and AES scores were suggested by Field (2018) and Williamson, Xi, and Breyer (2012). They suggested the amount of variance in human scores when compared with AES can be identified by using the criterion value of κ_q , $r \geq 0.70$. This means that a value ≥ 0.70 indicates a high level of agreement and reliability between raters and is considered the standard to achieve.

Analysis of Comparisons for Item #1

Item #1: “List the health history data (12 items) that would be helpful to collect (3 marks)” was scored and the following comparisons were calculated. Actual scores ranged from 0 to 3.00 and also included scores of 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, and 2.75.

Agreement measures and reliability coefficients. The agreement measures calculated for this study include exact percentage agreement and adjacent percentage agreement. The adjacent agreement percentage includes the exact agreement measure and scores that differ by 0.25 marks which equates to an 8.33 percent difference in agreement. Since item #1 is scored out of 3.00 marks with increments of 0.25, the closest adjacent score to measure is a difference in agreement of 0.25 marks. Cohen’s Kappa κ , Quadratic Weighted Kappa κ_q , and Pearson r , were calculated for the reliability coefficient measures. Table 4.1 outlines the results for item #1.

Table 4.1. Agreement Measures and Reliability Coefficients for Item #1.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.19	0.33	0.32
Exact + Adj % agreement**	0.35	0.47	0.39
Kappa	0.40	0.51	0.28
QWK	0.37	0.72	0.29
Pearson r	0.54	0.72	0.41

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and within 0.25 marks or 8.33 percent.

The exact agreements between human raters and AES ranged from 0.19 to 0.33 and the adjacent agreement measures between human raters and AES ranged from 0.35 to 0.47. The agreement measures between HR1 and AES were the highest and the agreement measures between the two human raters were the lowest. The degree of disagreement between human raters raises some concerns about rater preparedness and subjectivity in scoring. Even with a detailed scoring rubric, the human raters lacked agreement on scoring this item.

The reliability coefficients outlined in Table 4.1 indicate that only the reliability coefficient values for HR1 compared with AES met the criterion value of ≥ 0.70 . The reliability coefficient values for the other comparisons (0.54, 0.41, 0.37, 0.29) were below this criterion value meaning that the reliability between HR1 and AES was more consistent than the other comparisons. See Figure 4.10 for a visual representation of these results.

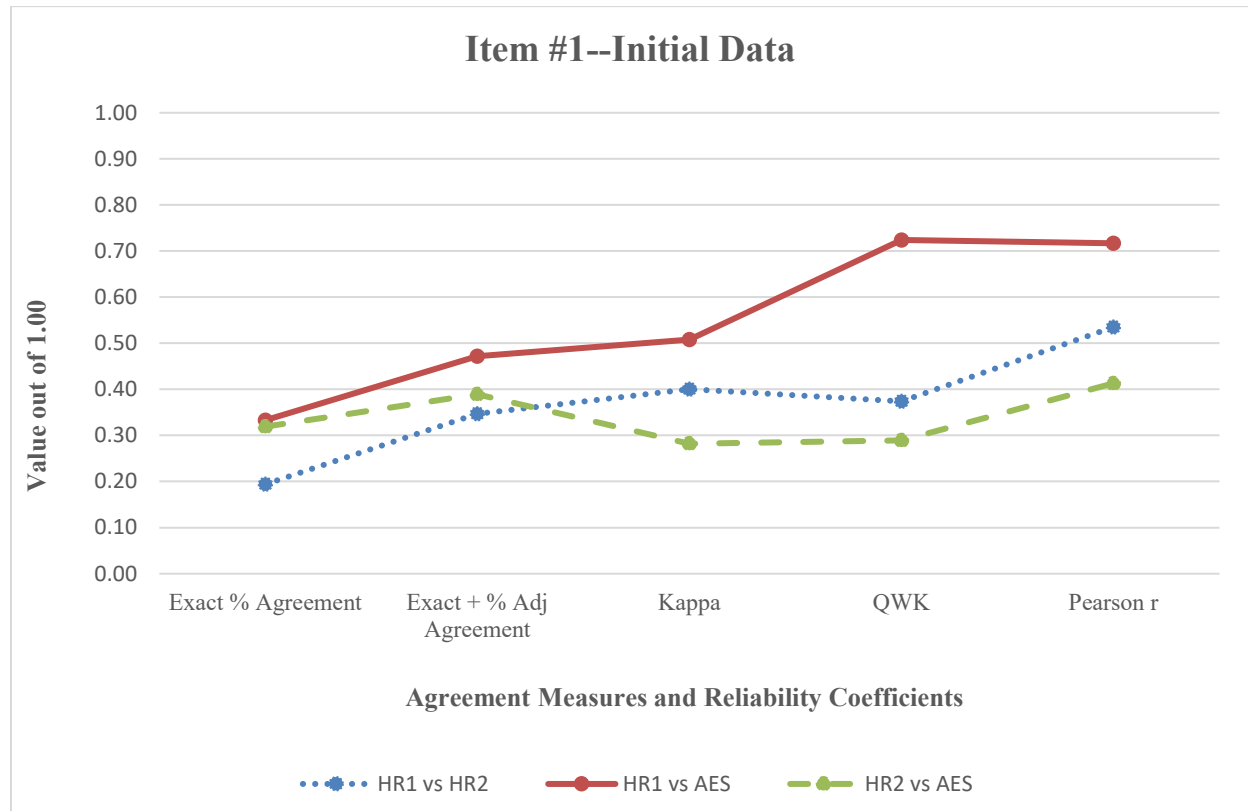


Figure 4.10. Agreement Measures and Reliability Coefficients for Item #1.

Score discrepancy analysis. Score discrepancy analysis gives a detailed overview of the differences in the overall scores. It is important to include this analysis in the study to determine the percentage of scores that had high discrepancies. For example, a score discrepancy of ± 1.50 (out of a possible 3.00 marks) for item #1 means that the difference between the raters was 50 percent. This obviously has a major impact on a student’s overall score.

To calculate score discrepancy measures, the absolute score differences were calculated within Excel for the following equations and the results are outlined in Table 4.2:

$$\text{Score}_{HR1} - \text{Score}_{HR2}$$

$$\text{Score}_{HR1} - \text{Score}_{AES}$$

$$\text{Score}_{HR2} - \text{Score}_{AES}$$

Table 4.2. Score Discrepancy Analysis for Item #1.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
±0	0.19	0.33	0.32
± 0.25 (8.33%)	0.15	0.14	0.07
± 0.50 (16.67%)	0.28	0.36	0.17
± 0.75 (25%)	0.08	0.11	0.06
± 1.00 (33%)	0.17	0.01	0.24
± 1.25 (41.67%)	0.06	0.03	0.07
≥ ± 1.50 (50%)	0.07	0.01	0.08

The results in Table 4.2 outline an important aspect of human and machine scoring which is the impact on student scores. When looking at the score discrepancy for HR1 and AES, over 83 percent of the student scores varied less than 17 percent. When considering a difference between scores of up to 25 percent, 94 percent of the student scores by HR1 and AES achieved this. Conversely, only 71 percent of the student scores by both human raters varied by less than 25 percent and over 12 percent of the student scores varied by almost 50 percent.

When analyzing the results from the score discrepancy results, it is essential to note the disparity and impact on overall student scores. In the graph below (see Figure 4.11), the large disparity between scores and its impact on students' scores in the range of 17 percent to 33 percent difference is clearly visualized when comparing HR1 and HR2. Whereas the disparity and impact on students' scores when comparing HR1 and AES is clearly less which results in more equity in students' scores. Overall, the agreement measures and reliability scores were highest for comparisons between HR1 and AES.

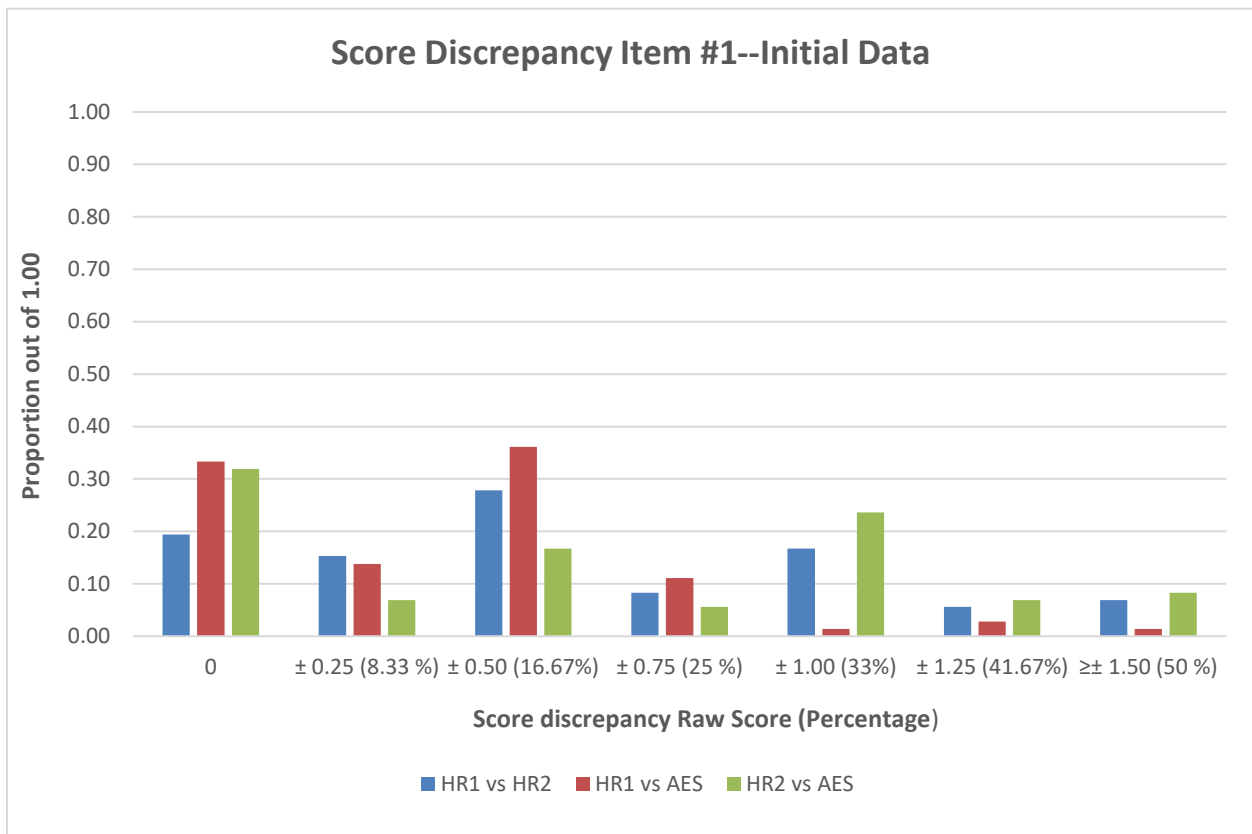


Figure 4.11. Score Discrepancy for Item #1.

Analysis of Comparisons for Item #2

Item #2: “Identify which physical assessments (8 items) you should perform (2 marks)” was scored and the following comparisons were calculated. Actual scores ranged from 0 to 2.00 and also included scores of 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, and 1.75.

Agreement measures and reliability coefficients. The agreement measures calculated for this item include exact percentage agreement and adjacent percentage agreement. The adjacent agreement percentage includes the exact agreement measure and scores that differ by 0.25 marks which equates to a 12.50 percent difference in agreement. Since item #2 is scored

out of 2.00 marks with increments of 0.25, the closest adjacent score to measure is a difference in agreement of 0.25 marks. The reliability coefficients of Cohen's Kappa κ , Quadratic Weighted Kappa κ_q , and Pearson r , were also calculated for item #2. Table 4.3 outlines the results for item #2.

Table 4.3. Agreement Measures and Reliability Coefficients for Item #2.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.40	0.54	0.40
Exact + Adj % agreement**	0.56	0.72	0.54
Kappa	0.11	0.39	0.16
QWK	0.57	0.76	0.39
Pearson r	0.62	0.75	0.49

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and within 0.25 marks or 12.50 percent.

The exact agreements between human raters and AES ranged from 0.40 and 0.54 and the adjacent agreement measures between human raters and AES ranged from 0.54 and 0.72. The agreement measures between HR1 and AES were the highest and the agreement measure between the human raters and HR2 with AES were the lowest.

As noted in the results for comparisons in the analysis for item #1, reliability coefficients, ≥ 0.70 indicate the criterion standard for reliability. As outlined in the Table 4.3, only the reliability coefficient values for HR1 compared with AES met this criterion value of ≥ 0.70 . The reliability coefficient values for the other comparisons (0.62, 0.49, 0.57, 0.39) were below this criterion value meaning that the reliability between HR1 and AES was more

consistent than the other comparisons. See Figure 4.12 for a visual representation of these results.

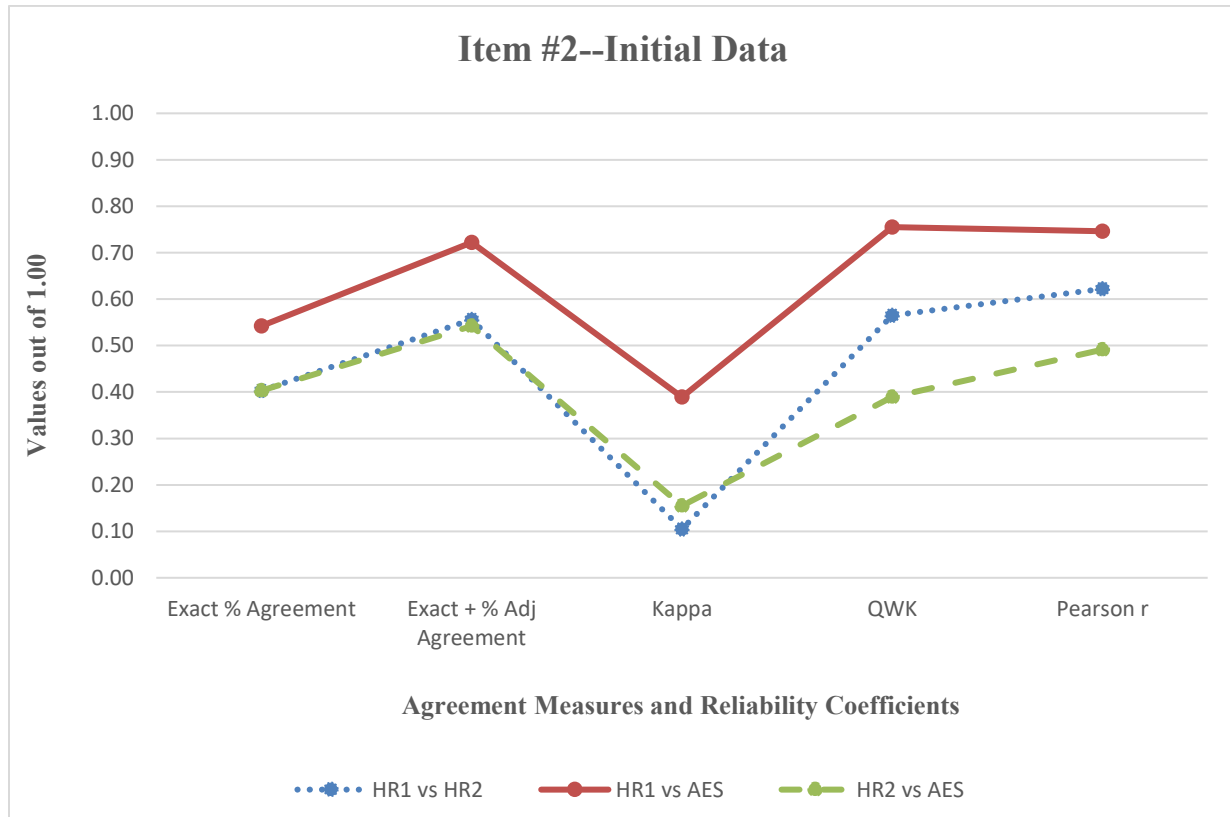


Figure 4.12. Agreement Measures and Reliability Coefficients for Item #2.

Score discrepancy analysis. As previously noted, score discrepancy analysis gives a detailed overview of the differences in the overall scores. Since item #2 was scored out of a possible total of 2.00 marks with 0.25 score increments, the differences have more impact on the total student score. The score discrepancy measures were calculated and are outlined in Table 4.4.

Table 4.4. Score Discrepancy Analysis for Item #2.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
±0	0.40	0.54	0.40
± 0.25 (12.50%)	0.15	0.18	0.14
± 0.50 (25%)	0.33	0.25	0.26
± 0.75 (37.50%)	0.07	0.01	0.06
≥± 1.00 (50%)	0.04	0.01	0.14

The results in Table 4.4 outline the impact on student scores. When looking at the score discrepancy for HR1 and AES, over 72 percent of the student scores varied less than 13 percent. When considering a difference between scores of up to 25 percent, 97 percent of the student scores by HR1 and AES achieved this. Less than 3 percent of the student scores varied by more than 37.50 percent between HR1 and AES. Conversely, only 55 percent of the student scores by both human raters varied by less than 13 percent and over 11 percent of the student scores varied by more than 37.50 percent. This means that score discrepancy between human raters was higher than the discrepancy between HR1 and AES. Overall, the agreement between HR1 and AES was the highest.

In the graph below (see Figure 4.13), the disparity between scores and the impact on students' scores in the 25 percent range is clearly visualized when comparing HR1 and HR2. Whereas the disparity and impact on students' scores when comparing HR1 and AES is clearly less which results in more equity in students' scores.

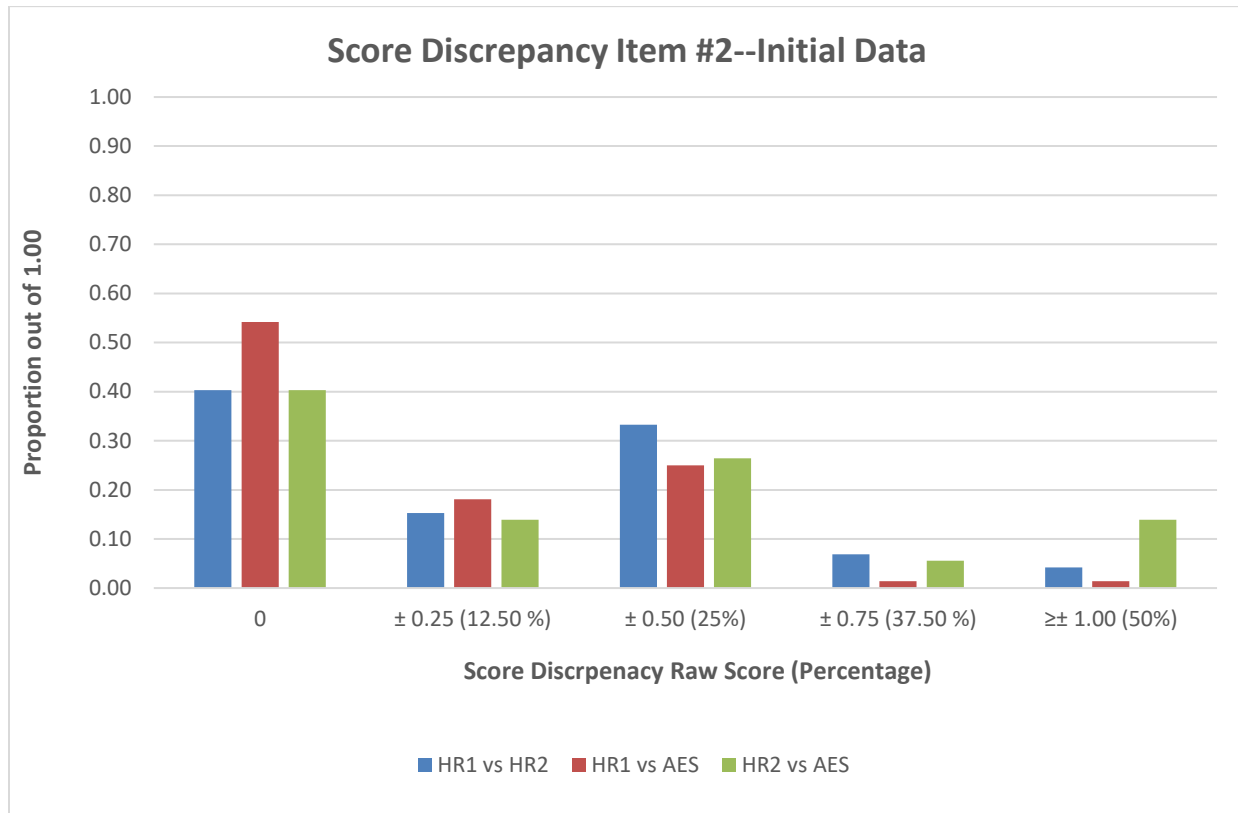


Figure 4.13. Score Discrepancy for Item #2.

Analysis of Comparisons for Item #3

Item #3: “List possible diagnostic tests that should be prescribed for Mrs. S (1 mark)” was scored and the following comparisons were calculated. Actual scores ranged from 0 to 1.00 and also included scores of 0.50.

Agreement measures and reliability coefficients. The agreement measures calculated for this item include exact percentage agreement and adjacent percentage agreement. The adjacent agreement percentage includes the exact agreement measure and scores that differ by 0.50 marks which equates to a 50 percent difference in agreement. Since item #3 is scored out of 1.00 mark with increments of 0.50, the closest adjacent score to measure is a difference in agreement of 0.50 marks. The reliability coefficients of Cohen’s Kappa κ , Quadratic Weighted

Kappa κ_q , and Pearson r , were also calculated for item #3. Table 4.5 outlines the results for item #3.

Table 4.5. Agreement Measures and Reliability Coefficients for Item #3.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	1.00	0.96	0.96
Exact + Adj % agreement**	1.00	1.00	1.00
Kappa	0.93	0.61	0.61
QWK	1.00	0.71	0.71
Pearson r	0.94	0.81	0.81

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and within 0.50 marks or 50 percent.

The exact agreements between human raters and AES were 0.96 meaning that almost 96 percent of the scores were scored exactly the same and the adjacent agreement measures meeting the 100 percent standard. It is essential to note that the scores between human raters were perfectly matched for this question whereas the machine scoring and human raters were not perfectly matched. The results for the agreement measures and reliability coefficients are identical because the scores by the human raters were identical for the entire data set. Item #3 is the only item to have higher reliability and agreement measures between human raters than the other comparisons. This may be related to the fact that the correct responses in item #3 are standard responses requiring very little interpretation and subjectivity. For example, the answers MRI, CT, bloodwork, and ECG are correct. There is very little variation in these responses which make this simpler for humans to agree on the scores. AES had to learn these responses

through keywords and the training set and then determine that the students had to put at least 2 of these keywords in their answers. With a larger data set, AES may achieve the same agreement measures as human raters.

As noted in the results for comparisons in the analysis for item #1, reliability coefficients, ≥ 0.70 indicate the criterion standard for reliability. As outlined in the Table 4.5, the human raters achieved reliability coefficients of 1.00 and the human raters with AES achieved reliability coefficients of 0.96 which is almost perfectly matched. The reliability coefficient values for the all comparisons were well above the ≥ 0.70 criterion standard for reliability.

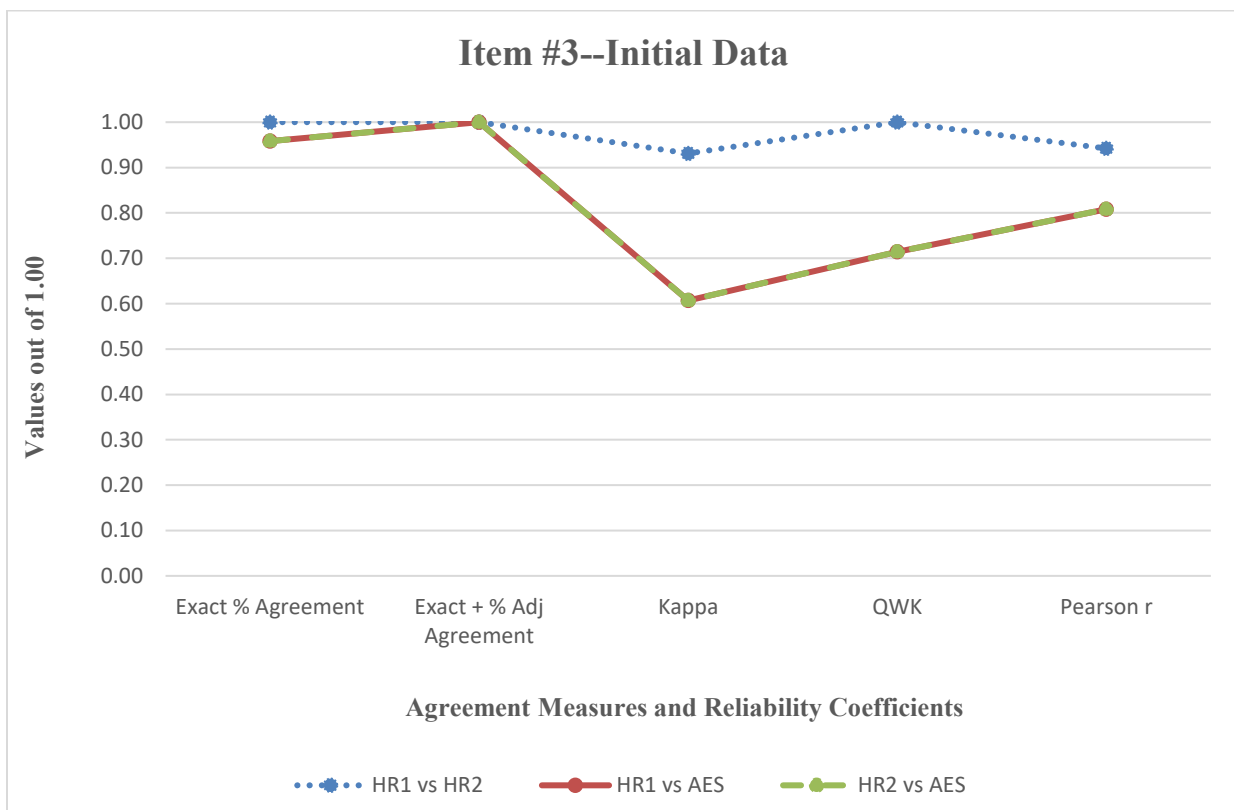


Figure 4.14. Agreement Measures and Reliability Coefficients for Item #3.

Score discrepancy analysis. Since item #3 was scored out of a possible total of 1 marks with 0.50 score increments, the differences have more impact on the total student score meaning that the impact on a student's score could be 50 percent difference. The score differences were calculated and the results are outlined in Table 4.6:

Table 4.6. Score Discrepancy Analysis for Item #3.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
±0	1.00	0.96	0.96
≥ ± 0.50 (50%)	0	0.04	0.04

Overall, the score discrepancy was minimal between AES and humans with only a few instances of score discrepancy being at a difference of 0.50. However, the impact of those few instances on those students results in a 50 percent difference in score. Being that this question is out of 1.00 mark, the overall impact on the exam total is minimal. The score discrepancy between human raters was 0 which means that all responses scored by the human raters were exactly the same. These results are illustrated in Figure 4.15.

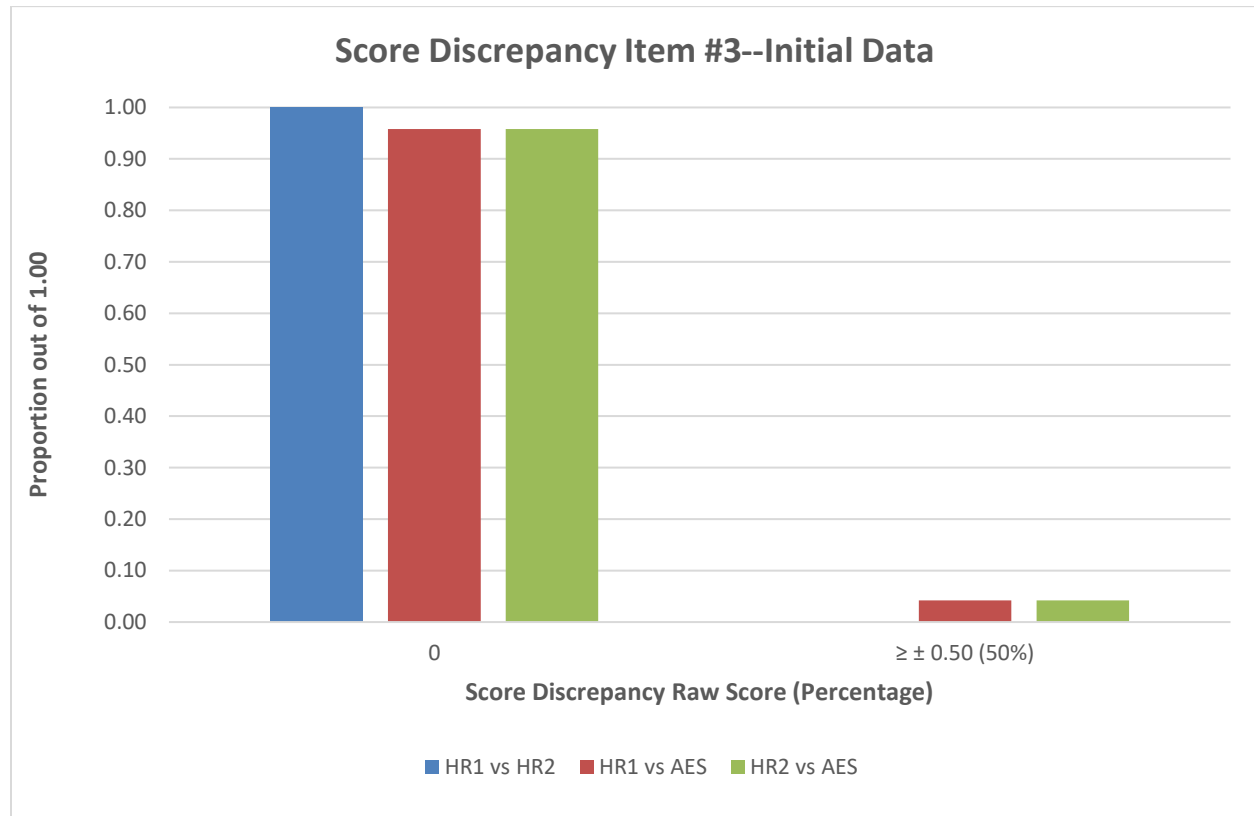


Figure 4.15. Score Discrepancy for Item #3.

Analysis of Comparisons for Item #4

Item #4: “Identify possible nursing interventions (8 items) to help Mrs. S (4 marks)” was scored and the following comparisons were calculated. Actual scores ranged from 0 to 4.00 and also included scores of 0.50, 1.00, 1.50, 2.00, 2.50, 3.00, and 3.50.

Agreement measures and reliability coefficients. The agreement measures calculated for this item include exact percentage agreement and adjacent percentage agreement. The adjacent agreement percentage includes the exact agreement measure and scores that differ by 0.50 marks which equates to a 12.50 percent difference in agreement. Since item #4 is scored out of 4.00 marks with increments of 0.50, the closest adjacent score to measure is a difference in agreement of 0.50 marks. The reliability coefficients of Cohen’s Kappa κ , Quadratic Weighted

Kappa κ_q , and Pearson r , were also calculated for item #4. Table 4.7 outlines the results for item #4.

Table 4.7. Agreement Measures and Reliability Coefficients for Item #4.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.15	0.43	0.13
Exact + Adj % agreement**	0.42	0.81	0.33
Kappa	0.03	0.28	0.03
QWK	0.25	0.68	0.17
Pearson r	0.59	0.68	0.39

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and within 0.50 marks or 12.50 percent.

The exact agreements between human raters and AES ranged from 0.13 and 0.43 and the adjacent agreement measures between human raters and AES ranged from 0.33 and 0.81.

Overall, the agreement measures between HR1 and AES were the highest and the agreement measure between the human raters and HR2 with AES were the lowest. An important observation is that the agreement between HR1 and AES was over 80 percent for scores within 12.50 percent.

As noted in the results for comparisons in the analysis for item #4, reliability coefficients, ≥ 0.70 indicate the criterion standard for reliability. As outlined in the Table 4.7, only the reliability coefficient values for HR1 compared with AES (0.68 and 0.68) were close to meeting this criterion value of ≥ 0.70 . The reliability coefficient values for the other comparisons (0.59,

0.39, 0.25, 0.17) were well below this criterion value meaning that the reliability between HR1 and AES was more consistent than the other comparisons.

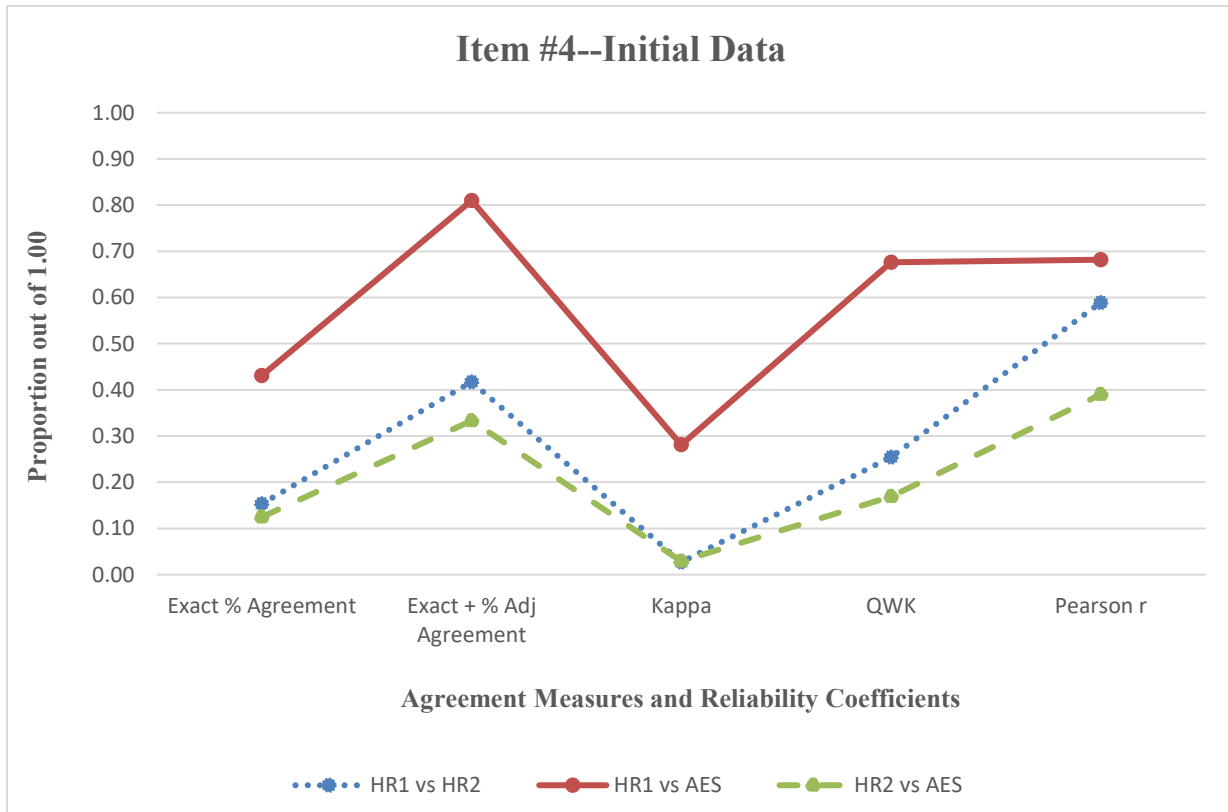


Figure 4.16. Agreement Measures and Reliability Coefficients for Item #4.

Score discrepancy analysis. Since item #4 was scored out of a possible total of 4.00 marks with 0.50 score increments, the differences impact the total student score. The score discrepancy measures were calculated and the results are outlined in Table 4.8.

Table 4.8. Score Discrepancy Analysis for Item #4.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
±0	0.15	0.43	0.13
± 0.50 (12.50%)	0.26	0.38	0.21
± 1.00 (25%)	0.18	0.18	0.26
± 1.50 (37.50%)	0.29	0.01	0.28
± 2.00 (50%)	0.11	0	0.11
± 2.50 (62.50%)	0	0	0.01
≥± 3.00 (75%)	0	0	0

When looking at the score discrepancy for HR1 and AES, over 81 percent of the student scores varied less than 13 percent. When considering a difference between scores of up to 25 percent, almost 99 percent of the student scores by HR1 and AES achieved this. Less than 2 percent of the student scores varied by more than 37.50 percent between HR1 and AES.

Conversely, only 42 percent of the student scores by both human raters varied by less than 13 percent and over 11 percent of the student scores varied by more than 37.50 percent. This means that score discrepancy between human raters was higher than the discrepancy between HR1 and AES.

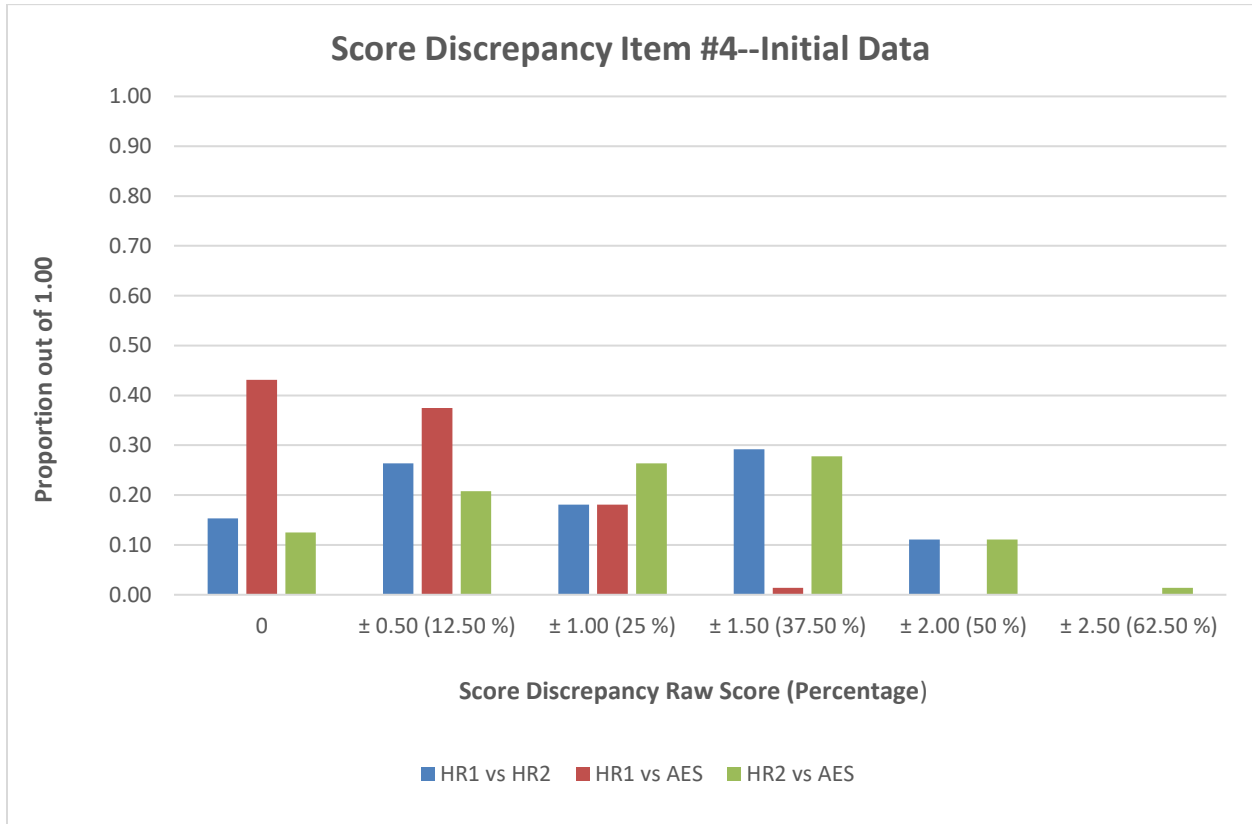


Figure 4.17. Score Discrepancy for Item #4.

The results from the analysis and comparisons for the initial 359 responses to 4 items show that there is higher reliability between HR1 with AES for items #1, #2, and #4 when compared with reliability coefficients between HR1 with HR2 and HR2 with AES. Item #3 showed high reliability in all comparisons but also showed perfect reliability between both human raters. These overall results are summarized in section 5 of this chapter.

Applying the AES Model

Once the AES model was developed and the results were analyzed for the items used in this study, the opportunity to apply the model to additional responses was identified. Section

Four is the overview of the results of analysis and comparisons of the AES model applied to newly gathered data.

Section Four: Results from Analysis of Application of AES Model

As an addition to the originally proposed study, the opportunity to apply the AES model that was developed for the initial 359 student responses was identified and the plan to apply the AES model to score a new set of student responses was implemented. In June, 2018, 40 nursing students at the Faculty of Nursing, University of Alberta, completed the exact four items as part of the completion of a 300 level nursing course. The examination was in similar conditions as the original examinations in April, 2016 and was run on the same *eclass* platform for the University of Alberta course delivery system.

A request to amend the original study was submitted and approved by the Research Ethics Board at the University of Alberta Research Ethics Office to include the collection and analysis of additional data (see Appendix H). Following ethics approval, permission to access the data was requested and granted by Dr. Olive Yonge, Vice Dean, Faculty of Nursing, University of Alberta (see Appendix I). Once ethics approval and permission to access the data were granted, the processes for data export were initiated following the exact same procedures as the export of the initial data.

Categories for Performance

The 300 level nursing course is focused on application of classroom content to the clinical setting. The overall course outcomes are focused on integrating previously learned concepts in patient care and nursing practice to a clinical setting including demonstration of critical thinking, clinical judgement, and decision making skills for patient care. Students

receive evaluations for clinical practice and course content. Many of the course concepts are covered using patient scenarios and evaluations are based on observed performance in clinical settings as well as written assessments.

In order to communicate meaningful evaluative criteria to nursing students, categories are used to describe levels of performance. Score categories are often used in educational measurement to ensure the students understand their score (Baldwin, Fowles, & Livingston, 2008; Kuo, Chen, Yang, & Mok, 2016; Minnich et al, 2018). The following categories are commonly used for evaluation in nursing education. 1) Unacceptable—the student is not meeting the standards for safe patient care or nursing practice; 2) Developing—the student is demonstrating beginning knowledge and performance for safe patient care; 3) Competent—the student is demonstrating understanding and performance of safe nursing practice and patient care; 4) Proficient—the student is demonstrating nursing practice and patient care that is above the minimum safe practice standards; and 5) Excellent—the student is demonstrating higher than expected safe nursing practice and patient care for the level of the course. To reflect these 5 categories of student achievement, parallel categories were developed for the scoring of the four items and are outlined in the results sections for each item. These categories of performance were approved by the teaching team leads (Dr. Simon Palfreyman and Ms. Karen Sylte) and determined to be reflective of the content included in the items.

All the original procedures for data export, removal of student identification, preprocessing, creating the “dummy class” for scoring by the second human rater, uploading data, use of keywords, and measures were followed for the data for the application of AES. The amount of time required by HR1 to score the 40 student responses to 4 items was 3.33 hours which averages 5 minutes per student. The amount of time required by HR2 to score the 40

student responses to 4 items was 4 hours which averages 6 minutes per student. The total amount of time required by AES to score all 4 items for the 40 student responses was less than one minute which averages just over 1 second per student.

Once the data was scored by both human raters and AES, the same analyses were completed for the four items using the same agreement measures, reliability coefficients, and score discrepancy analyses. Categories for student achievement were followed as outlined.

Analysis of Comparisons for Item #1—Application of AES Model

Item #1 was scored using the following five score categories:

- 1) Unacceptable—score of 0, 0.25, or 0.50
- 2) Developing—score of 0.75, 1.00, or 1.25
- 3) Competent—score of 1.50, 1.75, or 2.00
- 4) Proficient—score of 2.25 or 2.50
- 5) Excellent—score of 2.75 or 3.00.

Agreement measures and reliability coefficients. The agreement measures and reliability coefficients calculated for the data from the application of AES model for item #1 are outlined in Table 4.9. As noted in previous sections, the standard of ≥ 0.70 for reliability coefficients and 75 percent for agreement measures were used.

Table 4.9. Agreement Measures and Reliability Coefficients for Item #1—Application of AES Model.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.80	0.60	0.58
Exact + Adj % agreement**	0.95	0.85	0.80
Kappa	0.21	0.20	0.07
QWK	0.49	0.40	0.16
Pearson r	0.52	0.56	0.28

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and ± 1 score.

Agreement measures for comparisons between human raters and human raters with AES were above the recommended standard of 75 percent agreement. When analyzing the results for this item, the human raters had the highest agreement with 95 percent agreement for exact plus adjacent percentage agreement measures. The measures for reliability coefficients were all below the recommended standard. These results are further discussed in the summary section for this item.

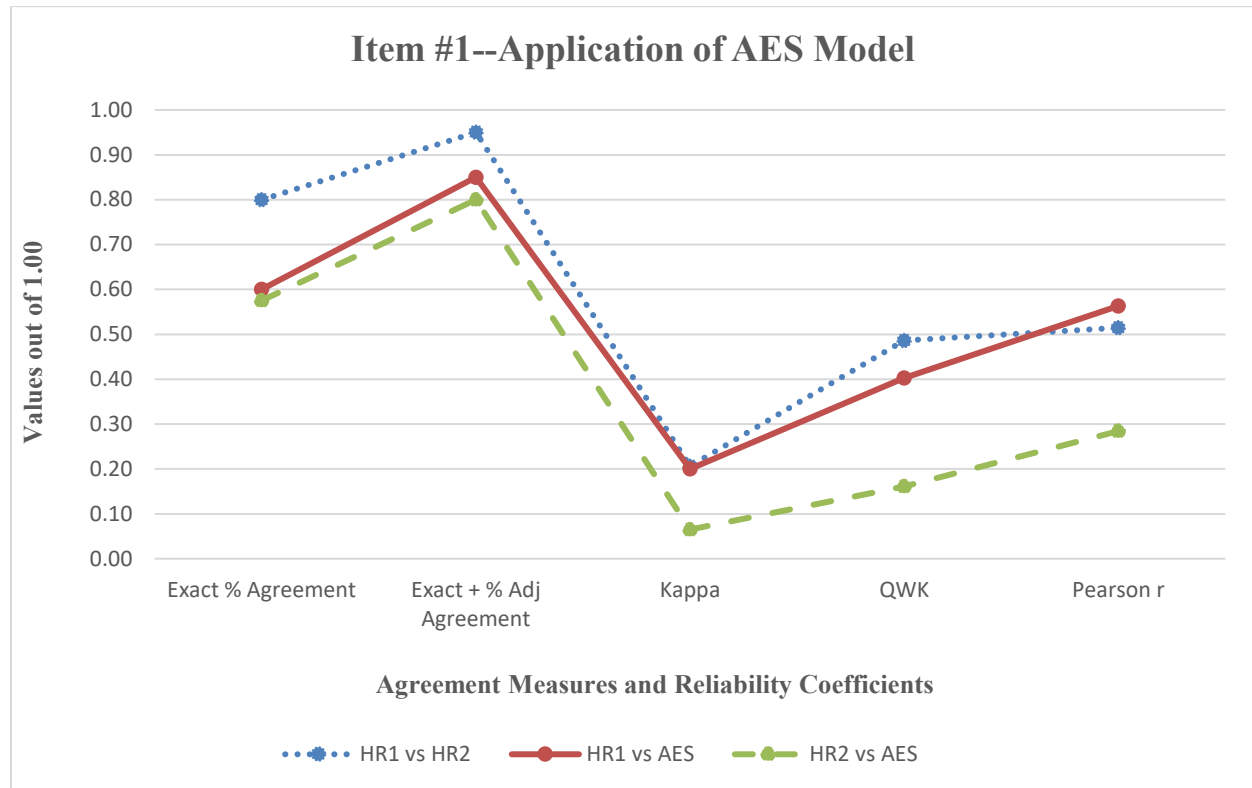


Figure 4.18. Agreement Measures and Reliability Coefficients for Item #1—Application of AES model.

Score discrepancy analysis. Score discrepancy analysis was completed for the data collected from the application of the AES model for item #1 and the results are outlined in Table 4.10. Score discrepancy was highest in the comparisons between HR2 and AES. Overall, the comparisons between human raters and HR1 with AES had the lowest score discrepancy which indicates higher agreement. All the comparisons met the recommended standard of over 75 percent agreement for ± 1 mark with HR1 and HR2 achieving a 95 percent agreement for this level. The results that are most concerning are the scores with a discrepancy of $\geq \pm 3$ marks. In the comparisons with HR1 and AES, two of the student scores differed by 3 or more score categories and in the comparisons with HR2 and AES, five of the student scores differed by 3 or

more score categories. This means there was almost no agreement on scoring these responses which is concerning for the impact on the students' scores and evaluation feedback.

Table 4.10. Score Discrepancy Analysis for Item #1—Application of AES Model.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
± 0	0.80	0.60	0.58
± 1	0.15	0.25	0.23
± 2	0.05	0.10	0.08
≥ ± 3	0	0.05	0.13

Figure 4.19 shows the score discrepancy for all comparisons. It is evident that although the score discrepancy values are low for over 80 percent of the scores, there are still many student scores that fall outside the level of acceptable agreement standards. This is further discussed in the summary section of this chapter.

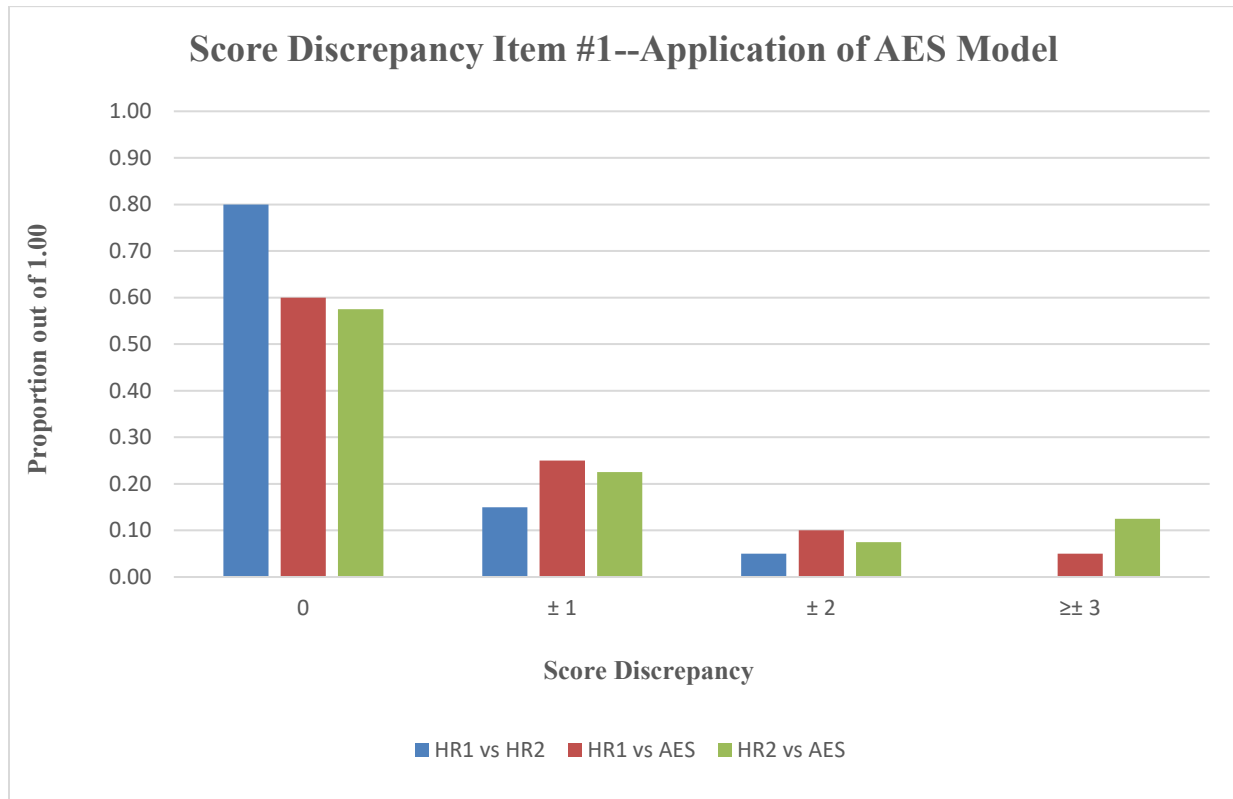


Figure 4.19. Score Discrepancy for Item #1—Application of AES Model.

Analysis of Comparisons for Item #2—Application of AES Model

Item #2 was scored using the following five score categories:

- 1) Unacceptable—score of 0 or 0.25
- 2) Developing—score of 0.50 or 0.75
- 3) Competent—score of 1.00 or 1.25
- 4) Proficient—score of 1.50 or 1.75
- 5) Excellent—score of 2.00.

Agreement measures and reliability coefficients. The agreement measures and reliability coefficients calculated for the application of AES model for item #2 are outlined in Table 4.11.

Table 4.11. Agreement Measures and Reliability Coefficients for Item #2—Application of AES Model.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.65	0.60	0.60
Exact + Adj % agreement**	0.95	0.98	0.98
Kappa	0.44	0.39	0.38
QWK	0.66	0.67	0.63
Pearson r	0.67	0.64	0.68

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and ± 1 score.

Agreement measures for all comparisons were well above recommended standards with exact plus adjacent percentage agreement results at 95 and 97.50 percent agreement. This indicates high levels of agreement between human raters and AES. The reliability coefficients for all comparisons were close to meeting the recommended standards. The results for Pearson r (0.67, 0.64, and 0.68) and Quadratic Weighted Kappa κ_q (0.66, 0.67, and 0.63) for all comparisons were close to meeting the recommended agreement standards but were not above 0.70. These results are visualized in Figure 4.20.

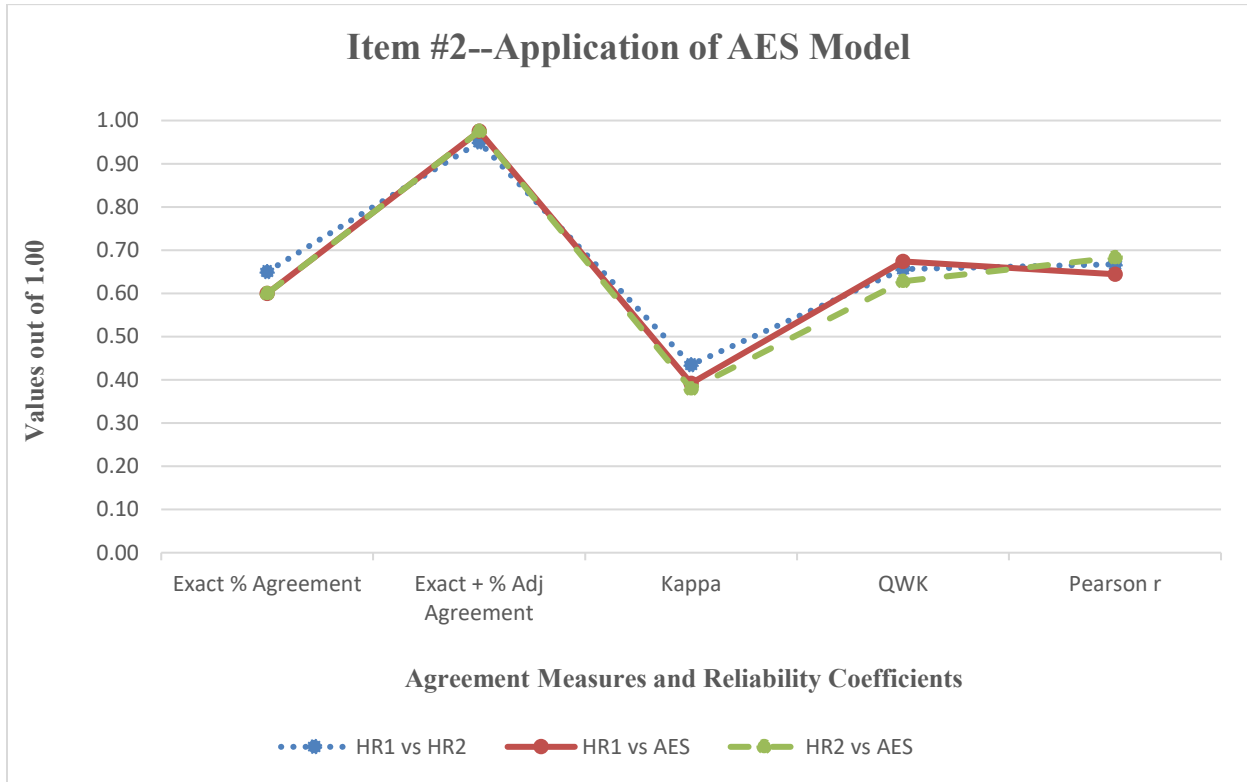


Figure 4.20. Agreement Measures and Reliability Coefficients for Item #2—Application of AES Model.

Score discrepancy analysis. Score discrepancy analysis was completed for the data collected from the application of the AES model for item #2 and the results are outlined in Table 4.12. All comparisons were well above the recommended agreement standard with over 95 percent agreement for ± 1 mark. None of the comparisons resulted in differences greater than ± 2 score categories and the score discrepancy values for ± 2 score categories were 0.05 and 0.03 which indicates very high level of agreement.

Table 4.12. Score Discrepancy Analysis for Item #2—Application of AES Model.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
± 0	0.65	0.60	0.60
± 1	0.30	0.38	0.38
± 2	0.05	0.03	0.03
$\geq \pm 3$	0	0	0

Overall, the score discrepancy results were low meaning that the scores were very similar and a high level of agreement was achieved. These results are represented in Figure 4.21.

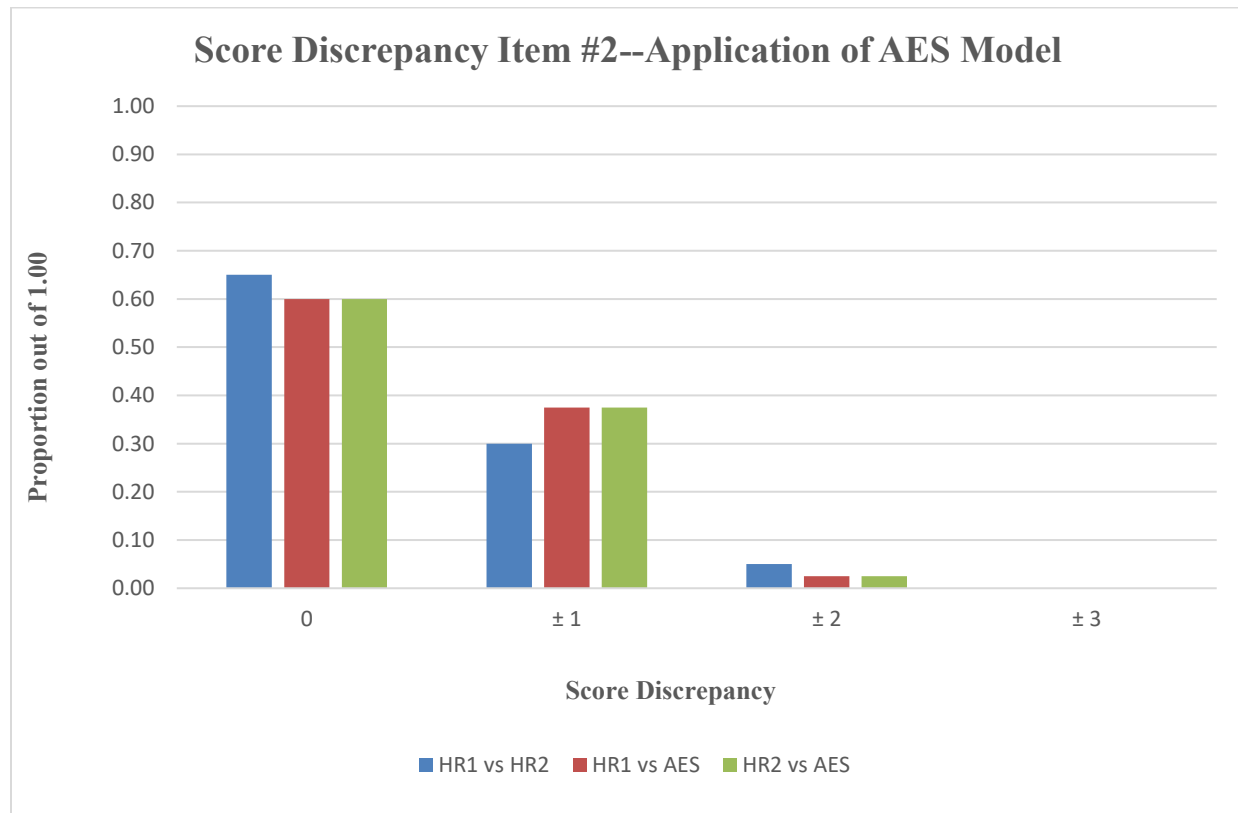


Figure 4.21. Score Discrepancy for Item #2—Application of AES Model.

Analysis of Comparisons for Item #3—Application of AES Model

Item #3 was scored using the following two score categories:

- 1) Unacceptable—score of 0
- 2) Proficient—score of 0.50 or 1.00.

Agreement measures and reliability coefficients. The agreement measures and reliability coefficients calculated for the application of AES model for item #3 are outlined in Table 4.13.

Table 4.13. Agreement Measures and Reliability Coefficients for Item #3—Application of AES Model.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	1.00	1.00	1.00
Exact + Adj % agreement**	1.00	1.00	1.00
Kappa	***	***	***
QWK	***	***	***
Pearson r	***	***	***

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and within 1 score.

***. No statistics are computed because all values are constant.

All comparisons between human raters and AES showed 100 percent agreement. Since the scores for all 40 responses received the exact same score, no statistics can be calculated for this item since all values are constant. It is important to note that the agreement measures meet the agreement standards and are well above the 75 percent agreement standard for all comparisons. These results are visualized in Figure 4.22.

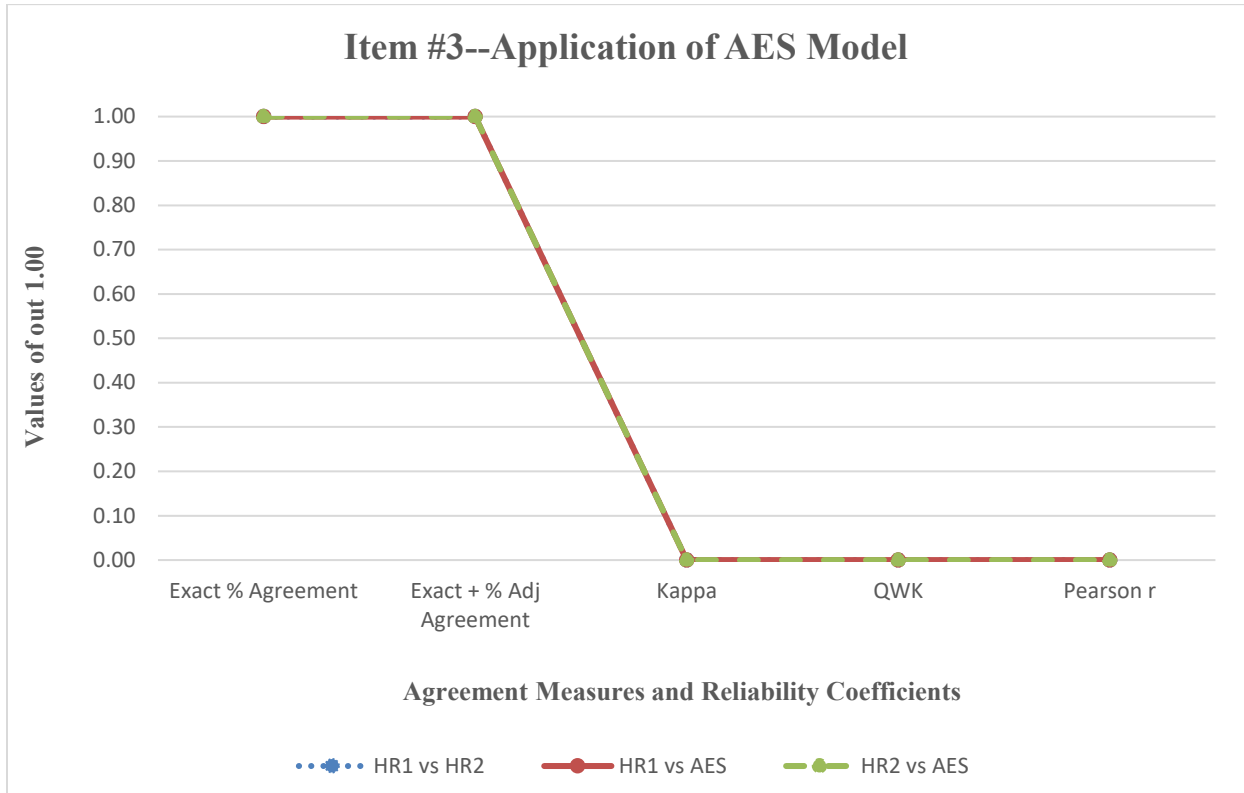


Figure 4.22. Agreement Measures and Reliability Coefficients for Item #3—Application of AES Model.

Score discrepancy analysis. Score discrepancy analysis was completed for the additional data collected from the application of AES model for item #3 and the results are outlined in Table 4.14 and visualized in Figure 4.23.

Table 4.14. Score Discrepancy Analysis for Item #3—Application of AES Model.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
± 0	1.00	1.00	1.00
≥ ± 1	0	0	0

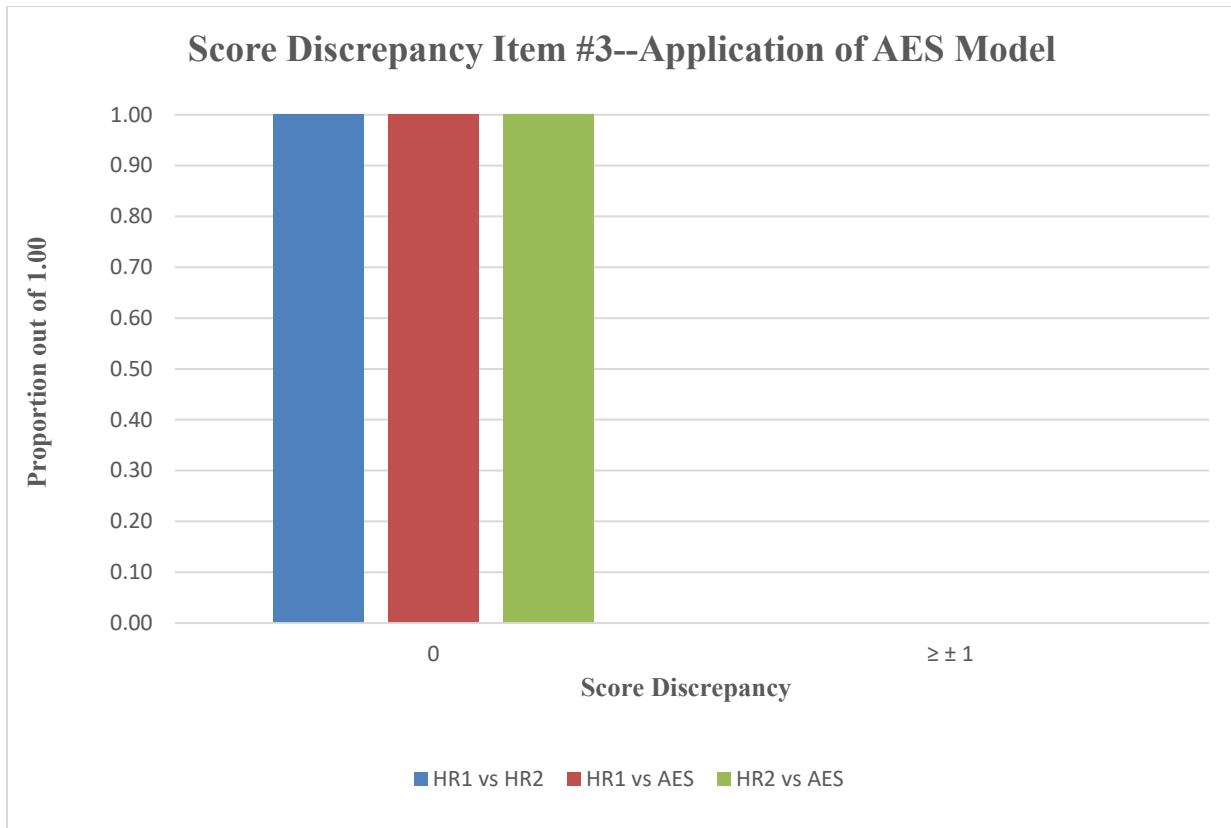


Figure 4.23. Score Discrepancy for Item #3—Application of AES Model.

Overall, all comparisons had no score discrepancy and indicate a perfect level of agreement of scores. All comparisons were clearly well above the recommended standards for agreement and reliability for this item.

Analysis of Comparisons for Item #4—Application of AES Model

Item #4 was scored using the following five score categories:

- 1) Unacceptable—score of 0
- 2) Developing—score of 0.50 or 1.00
- 3) Competent—score of 1.50 or 2.00
- 4) Proficient—score of 2.50 or 3.00

5) Excellent—score of 3.50 or 4.00.

Agreement measures and reliability coefficients. The agreement measures and reliability coefficients calculated from the application of the AES model for item #4 are outlined in Table 4.15.

Table 4.15. Agreement Measures and Reliability Coefficients for Item #4—Application of AES Model.

Measures	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % agreement*	0.55	0.55	0.18
Exact + Adj % agreement**	0.98	0.98	0.98
Kappa	0.37	0.16	0.04
QWK	0.66	0.21	0.11
Pearson r	0.68	0.68	0.22

*Exact % agreement is calculated by the number of times the scores are exactly the same.

** Exact + Adjacent % agreement is number of times the scores are exactly the same and ± 1 score.

The exact plus adjacent percentage agreement measures for all three comparisons was well above the recommended agreement standards at 97.50 percent agreement. HR2 compared with AES had a very low agreement measure for exact agreement at only 17.50 percent which resulted in the much lower reliability coefficient measures for HR2 and AES as indicated in Table 4.15. Overall, the agreement measures and reliability coefficients for human raters indicate high agreement. The results for HR1 with AES indicate very high agreement yet low reliability measures which may relate to the matrix computations used for these calculations. This is further discussed in the summary section of this chapter. These comparisons are shown in Figure 4.24.

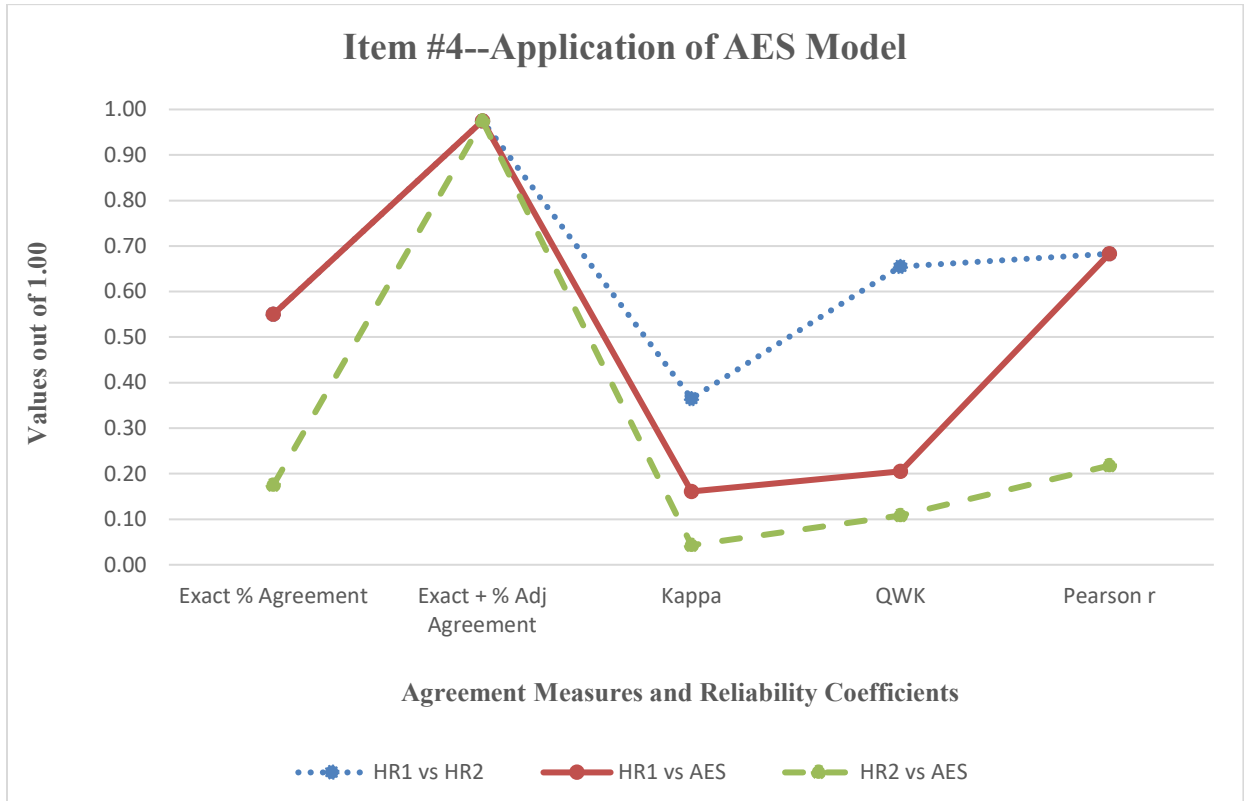


Figure 4.24. Agreement Measures and Reliability Coefficients for Item #4—Application of AES Model.

Score discrepancy analysis. Score discrepancy analysis was completed for the data from the application of AES model for item #4 and the results are outlined in Table 4.16.

Table 4.16. Score Discrepancy Analysis for Item #4—Application of AES Model.

Score Discrepancy	HR1 with HR2	HR1 with AES	HR2 with AES
± 0	0.55	0.55	0.18
± 1	0.43	0.43	0.80
± 2	0.03	0.03	0.03
± 3	0	0	0

The score discrepancy values are low for all three comparisons for scores within ± 1 and are well above the recommended level of agreement with 97.50 percent agreement. For scores ± 2 , only 1 student response in all comparisons differed by this amount which demonstrates an excellent level of agreement.

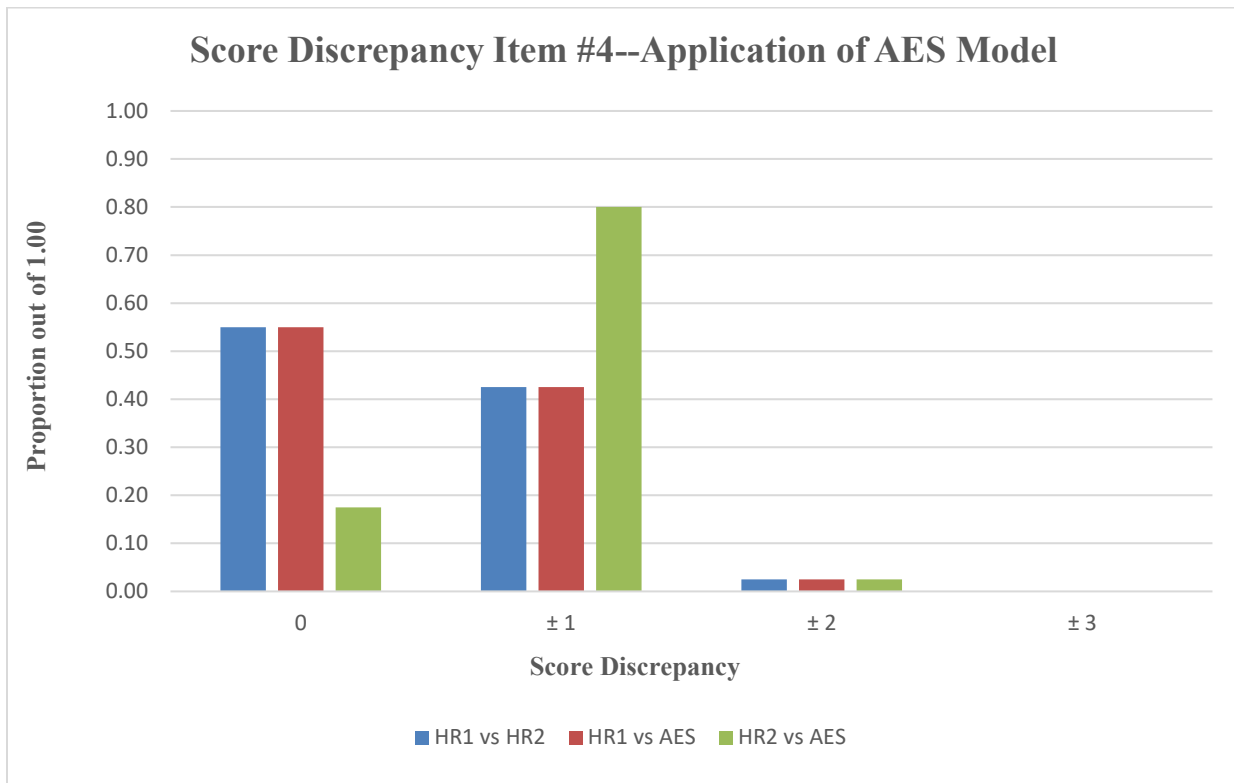


Figure 4.25. Score Discrepancy for Item #4—Application of AES Model.

Overall, the score discrepancy values are very low indicating a very high level of agreement for all comparisons for item #4.

Section Five: Summary and Discussion of Results

This section includes an overview of the comparisons for human raters, HR1 with AES, and HR2 with AES for all data collected in 2016 and 2018. Also included in this section is an

overall summary and discussion of the results for each item reflective of all data collected and analyzed.

Comparisons Between Human Raters

The results of the comparisons between human raters were very low in agreement and reliability measures for the initial data set collected in 2016, but much higher for the results from the application of the AES model. All results for items #1, #2, and #4 were well below the recommended standards indicating a concerning lack of agreement between human raters.

The results from the application of the AES model collected in 2018 indicate higher levels of agreement and reliability measures for items #1, #2, and #4 between human raters when considering the results from the initial data set (see Table 4.17). Item #3 achieved perfect agreement levels and reliability for human raters for the initial data and application of the AES model.

Table 4.17. Summary of Comparisons between Human Raters.

Measures	Item							
	1	2	3	4	1A	2A	3A	4A
Exact % Agreement	0.19	0.40	1.00	0.15	0.80	0.65	1.00	0.55
Exact + Adj %	0.35	0.56	1.00	0.42	0.95	0.95	1.00	0.98
Kappa	0.40	0.11	0.93	0.03	0.21	0.44	-	0.37
QWK	0.37	0.57	1.00	0.25	0.49	0.66	-	0.66
Pearson r	0.54	0.62	0.94	0.59	0.52	0.67	-	0.68

It is evident that the agreement and reliability measures for human raters were much higher for the results from the application of the AES model. Several factors may have contributed to these results. Two of the factors that are well supported in the literature are rater fatigue and “drift” (Almond, 2014; Tan, Kim, Paek, & Siang, 2009). The initial data set included

responses from 359 students which were scored by both human raters. The total number of responses that were scored was 1436 (359 x 4 items). The impact of fatigue and “drift” is increased when scoring this many responses when compared with the much smaller data set of 40 student responses that were used for the application of the AES model. Another factor that may have affected the lack of agreement and reliability of the scores for the initial data set is the impact of the scoring on student achievement. The scores completed by HR1 were actual scores that were included in student grades for their university course. The scores completed by HR2 were done strictly for the purposes of this research study and were never included or communicated to the students. Although this is difficult to determine, the differences in the consequences for scoring by HR1, as opposed to HR2, may have affected the amount of diligence given to the task of scoring the responses. Finally, as noted in the literature review in chapter 2, humans are, by nature, subjective in scoring and can be affected by general mood and previous exposure to student performance (Attali, Lewis, & Steier, 2012; Williamson, Xi, & Breyer, 2012). These factors may account for the overall low levels of agreement between human raters on the initial set of data.

Comparisons Between HR1 and AES

Interestingly, the agreement and reliability measures for HR1 compared with AES met the recommended standards for the initial data set for items #1, #2, and #3 and indicate a substantial agreement. Item #4 was close to 0.70 indicating a moderate level of agreement. Item #3 was scored at near perfect agreement, which was less than the level of agreement for human raters scoring this item in the initial set of data but well above the standard. The results are outlined in Table 4.18.

The results from the application of the AES model indicate near perfect agreement measures for the exact plus adjacent percentage values. All data collected for these comparisons indicate high to perfect levels of agreement. The reliability measures for HR1 and AES were much higher for the initial data set than the results from the application of the AES model with all reliability measures being below the recommended level of 0.70 for reliability. It is important to consider the implications of low reliability measures occurring even with high agreement measures. When calculating Kappa κ , and Quadratic Weighted Kappa κ_q , a matrix is used. When very high levels of agreement occur, the matrix has mostly “0”s throughout the matrix with a few “1”s scattered throughout. With almost no other values in the matrix, the calculation does not detect the levels of disagreement between values which results in low reliability values for Kappa κ and Quadratic Weighted Kappa κ_q (Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Field, 2018; Yarnold, 2016). This has been called the Paradox of Kappa (Yarnold, 2016) and is evident in the results from the application of the AES model in item #4. It is essential to consider these results in the context of students and the impact on the students’ scores.

Overall, comparisons between HR1 and AES were higher in agreement measures and lower in reliability measures on the data from the application of the AES model than the measures on the initial dataset. This may be related to the fact that the initial data set was used as the overall training set based upon the scores from HR1. The scoring model was developed based upon HR1 scores which may account for high levels of agreement between HR1 and AES and the low levels of agreement between AES and HR2.

Table 4.18. Summary of Comparisons between HR1 and AES.

Measures	Item							
	1	2	3	4	1A	2A	3A	4A
Exact % Agreement	0.33	0.54	0.96	0.43	0.60	0.60	1.00	0.55
Exact + Adj %	0.47	0.72	1.00	0.81	0.85	0.98	1.00	0.98
Kappa	0.51	0.39	0.61	0.28	0.20	0.39	-	0.16
QWK	0.72	0.76	0.71	0.68	0.40	0.67	-	0.21
Pearson r	0.72	0.75	0.81	0.68	0.56	0.64	-	0.68

Comparisons Between HR2 and AES

Overall, the comparison between HR2 and AES achieved lower levels of agreement and reliability measures when compared with human raters and HR1 with AES on the initial data set and slightly lower levels for reliability measures for the data from the application of the AES model. The results for this comparison for the data from the application of the AES model indicate much higher levels of agreement than the initial data set, which are outlined in Table 4.19.

Table 4.19. Summary of Comparisons between HR2 and AES.

Measures	Item							
	1	2	3	4	1A	2A	3A	4A
Exact % Agreement	0.32	0.40	0.96	0.13	0.58	0.60	1.00	0.18
Exact + Adj %	0.39	0.54	1.00	0.33	0.80	0.98	1.00	0.98
Kappa	0.28	0.16	0.61	0.03	0.07	0.38	-	0.04
QWK	0.29	0.39	0.71	0.17	0.16	0.63	-	0.11
Pearson r	0.41	0.49	0.81	0.39	0.28	0.68	-	0.22

This is possible due to the training set for AES being modeled from HR1 scores on the initial data set. It is evident that the agreement and reliability measures between HR1 and HR2

are low for the initial data set. Since the AES model was developed from the initial data set, it is possible this affected the comparisons between HR2 and AES. This will be further discussed in chapter 5.

The analysis of the comparisons between human raters and AES for the initial data set and application of the AES model data set indicate many positive results, which provide valuable information moving forward with the use of AES. Also, some challenges with the use of developing an AES model from an initial set of responses then applying it to new data sets for short essay constructed response items are identified. It is essential to consider the comparisons between human raters and AES, as well as the results for each item.

Summary of Analysis and Discussion for Item #1

The overall summary for the results for item #1 is outlined in Table 4.20. Item #1 had acceptable levels of agreement and reliability measures, but overall had the lowest level of agreement for all four items included in this study. When reviewing the item, many factors need to be considered, some of which relate to the design of the item, including expected responses. Item #1 has many possibilities that are considered correct. The scoring rubric includes 24 possible correct responses with many of these responses having multiple correct responses imbedded in them. For example, “PQRST” is one of several mnemonics used to remember the components of analyzing a symptom or sign. Other examples of mnemonics for the same content include: “PQRSTUV”, “OLDCARTS”, and “LATER SNAPS”. All of these mnemonics help learners remember the 8 to 10 components of sign or symptom analysis, which are included in a patient’s health history. It is essential to note that the scoring rubric also includes the 8 to 10 components that are reflected in the mnemonics. For example, the “T” in all of the mnemonics refers to “timing” of the sign or symptom. In this case, “timing” refers to the timing of the

patient's headache including onset and duration of the pain. "Onset", "duration", and "timing" are all included in the scoring rubric, which may make this item challenging to accurately score. This overlap of correct responses may lead to challenges with training the AES model and different interpretations between humans. A possible solution is to reduce the scope of the item to focus on fewer correct responses with little or no overlap.

Also related to the design of item #1 are the many versions of a correct response. For example, "family history" is included as the overall term for one part of the health history, but this can be implied by the response, "Does anyone in your family have diabetes?" It is challenging to determine if the student was competent in knowing the need to assess for the genetic conditions that are relevant in this situation (such as stroke, migraines, diabetes, heart disease, or thyroid disorders) and are critical to know about the family history of the patient. Or did the student just ask about diabetes and not consider the overall context of family history. The interpretation of these responses can be quite different and therefore challenging to score reliably. In future examinations and studies, it may be valuable to redesign item #1 to ensure higher quality of the item.

Table 4.20. Summary of Results for Item #1.

Measures	Item #1			Item #1A		
	HR1 with HR2	HR1 with AES	HR2 with AES	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % Agreement	0.19	0.33	0.32	0.80	0.60	0.58
Exact + Adj %	0.35	0.47	0.39	0.95	0.85	0.80
Kappa	0.40	0.51	0.28	0.21	0.20	0.07
QWK	0.37	0.72	0.29	0.49	0.40	0.16
Pearson r	0.54	0.72	0.41	0.52	0.56	0.28

Summary of Analysis and Discussion for Item #2

The overall summary for the results for item #2 is outlined in Table 4.21. Item #2 had high agreement and reliability measures for most comparisons and performed well on most measures. When compared to the analysis of item #1, item #2 seems to be better designed including the expected responses. There is minimal overlap in the correct responses and the correct responses are very objective in nature. For example, “handgrip” is one of the correct responses on the scoring rubric and in the list of keywords used. “Handgrip” is the name of one of the physical examination techniques used to assess for a stroke. Unlike item #1, there are few other ways to describe “handgrip” which reduces the amount of needed interpretation and subjectivity. Also, there is no other response imbedded in the response “handgrip” which reduces the overlap in the responses. Almost all of the correct responses have little to no room

for variances in interpretation and are focused on single concepts instead of multiple possibilities in one response.

Table 4.21. Summary of Results for Item #2.

Measures	Item #2			Item #2A		
	HR1 with HR2	HR1 with AES	HR2 with AES	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % Agreement	0.40	0.54	0.40	0.65	0.60	0.60
Exact + Adj %	0.56	0.72	0.54	0.95	0.98	0.98
Kappa	0.11	0.39	0.16	0.44	0.39	0.38
QWK	0.57	0.76	0.39	0.66	0.67	0.63
Pearson r	0.62	0.75	0.49	0.67	0.64	0.68

Summary of Analysis and Discussion for Item #3

The overall summary for the results for item #3 is outlined in Table 4.22. Item #3 had the highest agreement and reliability measures out of all four items. It is important to note that although item #3 achieved the highest levels of agreement, this item could be easily scored with other strategies that are much less complex than AES. Essentially, item #3 is a fill-in-the-blank item that could be scored with simpler software programs that are already available within the *eclass* platform used at the University of Alberta. The agreement measures for human raters and AES are considered perfect, but this same level of agreement could be achieved with more accessible and simpler software. Essentially, AES accurately scored this item when compared

with human raters and is an excellent strategy to implement for scoring this item, however, depending on the user, there may be simpler strategies to score this item as well as AES.

Table 4.22. Summary of Results for Item #3.

Measures	Item #3			Item #3A		
	HR1 with HR2	HR1 with AES	HR2 with AES	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % Agreement	1.00	0.96	0.96	1.00	1.00	1.00
Exact + Adj %	1.00	1.00	1.00	1.00	1.00	1.00
Kappa	0.93	0.61	0.61	-	-	-
QWK	1.00	0.71	0.71	-	-	-
Pearson r	0.94	0.81	0.81	-	-	-

Summary of Analysis and Discussion for Item #4

The overall summary for the results for item #4 is outlined in Table 4.23. Item #4 requires the student to analyze the situation and consider appropriate interventions to ensure patient safety and early recognition of a critical event.

Table 4.23. Summary of Results for Item #4.

Measures	Item #4			Item #4A		
	HR1 with HR2	HR1 with AES	HR2 with AES	HR1 with HR2	HR1 with AES	HR2 with AES
Exact % Agreement	0.15	0.43	0.13	0.55	0.55	0.18
Exact + Adj %	0.42	0.81	0.33	0.98	0.98	0.98
Kappa	0.03	0.28	0.03	0.37	0.16	0.04
QWK	0.25	0.68	0.17	0.66	0.21	0.11
Pearson r	0.59	0.68	0.39	0.68	0.68	0.22

Item #4 has fewer correct responses when compared with items #1 and #2. The overall agreement and reliability measures indicate a moderate to high level of agreement for most measures except some of the values for Kappa κ and Quadratic Weighted Kappa κ_q . This outcome was discussed previously in this summary section. Item #4 is similar to item #2 in that the correct responses have little overlap, relate to a single concept, and are less open to interpretation. For example, the response “start IV” has few different meanings—“start intravenous” or “initiate intravenous therapy”—all have the same meaning and require little interpretation. The design of item #4 including expected responses may have contributed to the high levels of agreement in the comparisons.

Conclusion

These results, analyses, related factors, identified challenges, and suggestions for moving forward provide valuable information for the assessment and measurement of higher-level

thinking skills in nursing education. Item development, rater preparedness, and implementation of AES to supplement scoring strategies are some of the areas needing further discussion and consideration. These will be included in the discussion in chapter 5.

Chapter 5: Discussion and Conclusion

Advances in technology continue to change the way we live and learn. Machine learning techniques are consistently advancing and shaping the way we connect with the world and learn through our experiences. These advances in technology and machine learning also impact how we assess learning and evaluate performance. Automated essay scoring (AES) is a developing technology that is increasingly recognized as a potential strategy for managing the challenges associated with testing and scoring written assessments (Dikli, 2006; Gierl et al, 2014; Kuo et al, 2016). In the United States of America, many large-scale educational assessments are using AES for evaluating written essays and gathering information (Latifi, 2016). As a result, development and research into the use of AES for written assessments is an important topic in educational measurement.

As the development of AES systems progresses and expands, the need to consider implementation of AES in multiple disciplines becomes essential. Health science professions use different language and terminology compared to other disciplines. Exploring the use of AES in the health sciences fields, specifically nursing, increases the information we have regarding the efficacy of AES overall. Several studies have been conducted using AES for scoring essay format items for assessing learning in medical education (Latifi, Gierl, & Boulais, 2013), but no published research in the area of using AES in assessing nursing education was found.

Health sciences education programs have been faced with finding solutions to manage the challenges of assessing higher-level thinking skills such as, critical thinking, clinical reasoning, and clinical judgement, in student learning. Accurate and comprehensive assessment of higher-level thinking skills is essential in nursing education to ensure the standards of safe patient care (Alfaro-LeFevre, 2017; Canadian Patient Safety Institute, 2008; Oermann & Gaberson, 2017). It

is well recognized that the use of examinations consisting of only selected-response items are insufficient to assess these higher-level thinking skills (Brookhart & Nikito, 2015; Kuo, Chen, Yang, & Mok, 2016; Oermann & Gaberson, 2017; Tankersley, 2007; Yang, Liu, & Morell, 2018). Also, there is always the possibility of successfully guessing the correct answer on a selected-response item without any knowledge of the content (Andrich & Marais, 2018; Kuo, Chen, Yang, & Mok, 2016; Sangwin & Jones, 2017). Although it is well recognized that constructed-response items are beneficial for assessing higher-level thinking skills (Andrich & Marais, 2018; Brookhart & Nikito, 2015; Kuo, Chen, Yang, & Mok, 2016; Oermann & Gaberson, 2017; Sangwin & Jones, 2016; Yang, Liu, & Morell, 2018) several challenges to the inclusion of constructed-response items in examinations exist. Constructed-response items are more costly and time-consuming to score, are affected by subjectivity and errors in scoring, and are susceptible to issues with reliability (Attali, Lewis, & Steier, 2013; Gierl et al, 2014; Kuo, Chen, Yang, & Mok, 2016; Shermis & Burstein, 2013). These challenges have resulted in the shift to exclusively using only selected-response items on exams—especially in large classes. By eliminating constructed-response items from learning assessments in nursing education, it is difficult to accurately assess whether nursing students are meeting the standards for safe nursing practice. It is imperative to find strategies to overcome the challenges associated with the inclusion of constructed-response items on examinations in nursing education programs.

The focus of this study was to explore the use of AES for assessing higher-level thinking skills in nursing education; specifically, to study the effectiveness of AES for scoring constructed-response items on nursing examinations and consider the potential uses of AES in nursing education assessments for the future.

This chapter is organized into three sections. The first section includes the purpose of the study, research questions, and an overview and summary of the research methods and results from this study. Section two includes a discussion of the limitations of the study. The third section includes suggestions for direction of future research.

Section One: Restatement of Research Questions and Summary of Methods and Results

The primary purpose of this research study was to investigate the effectiveness of using AES to assess higher-level thinking skills in nursing education. To facilitate this research, four constructed-response items were developed and included in nursing examinations. These items were based on an appropriate and realistic patient scenario reflective of a typical patient situation in a clinical setting for the level of students completing the examination.

Three research questions were addressed in this study:

- 1) Is AES as effective as human raters for scoring constructed-response items in nursing education in terms of accuracy and reliability?
- 2) Does AES score constructed-response items more efficiently than human raters?
- 3) Is AES a potential solution to overcome the challenges of time, cost, and subjectivity in scoring constructed-response items?

To answer these research questions, an analytical approach to compare differences between humans and AES in scoring four constructed-response items on nursing education examinations was initiated. An anonymized dataset was used.

In order to facilitate the implementation of AES and comparison analyses, it was essential to administer the items using a computer-based platform. At the time this research study was in planning stages, the only available process for including constructed-response items on examinations at the U of A involved the students writing the responses by hand in an

examination booklet. These examination booklets are 6 pages long and have been used for decades at the U of A for students to respond to essay-type items on examinations. It was imperative to construct a computer-based platform through the learning management system that would allow the students to key their responses to the constructed-response items rather than handwriting their answers. If this was not possible, all student responses would have required transcription which would have been costly and time-consuming. This obviously conflicts with the overall goal of this research study.

The researcher together with the Teaching and Learning Technology (TLT) team at the Faculty of Nursing (FON), U of A, developed the platform to administer the constructed-response items through the learning management system currently used at the U of A. In the examinations used for this study, students completed several selected-response items and the four constructed-response items within the same examination platform. The developed platform eliminated the need for students to complete their responses on a separate word document or unique file for the examination. All items were seamlessly included in the same format.

It is important to note that the platform that was developed for students to complete these four constructed-response items is now widely used to administer constructed-response items on examinations in multiple courses at the FON. As a result, the FON is one of the only faculties at the U of A that has eliminated the use of examination writing booklets and the practice of students writing responses to examination items by hand. All examinations in the FON are now completely computer-based, including constructed-response items, and students key in their responses to all items.

The 4 constructed-response items were administered to 359 nursing students in 2016 and were scored. All student responses were scored by one human rater (HR1) and recorded and

calculated into the students' grades. This process required 34 hours in total for HR1 to complete the scoring for 359 students. After the scores were recorded, the files were uploaded to excel and all student identifying data were removed.

In 2017, ethics approval was granted to access the scores and items from the examinations administered in 2016. The scored responses by HR1 were exported into appropriate file formats and used to develop an AES model for scoring. This process is described in detail in chapter 4.

In order to analyze the scores and compare AES with traditional scoring methods, a second human rater (HR2) was involved to score all four items for 359 students. This required 39 hours for HR2 to complete the scoring for all 359 students. Once the AES model was developed from the scores from HR1, the scores from HR1, HR2, and AES were compared and analyzed using agreement measures and reliability coefficients.

Agreement measures and reliability coefficients were computed and analyzed to study the differences in scoring between human raters and AES. Initially, the research plan was to use 40 percent of the dataset as the training set, however, the development of the AES model required larger numbers of data to accurately train the program. This resulted in needing to use 80 percent of the dataset for the training set, which meant lower numbers of responses being scored by AES. Once the AES model was developed and compared with HR1 and HR2, it was applied to a new dataset.

In 2018, 40 nursing students completed the same 4 constructed-response items as part of a 300-level nursing examination. The responses on the examination were again scored by HR1 and the scores were calculated into the students' grades. Student identifying data was removed to ensure anonymity of the responses. An amendment to the original research proposal was

submitted to the ethics committee and approval to access the responses and scores was granted. The AES model that was developed from the initial 359 student responses was applied to score the responses for 40 nursing students. To ensure similar conditions for comparisons, the same HR2 scored all 40 student responses. Agreement measures and reliability coefficients were calculated for the comparisons between AES, HR1, and HR2.

Summary of Results

The primary purpose of this research study was to investigate the effectiveness of using AES to assess higher-level thinking skills in nursing education. A secondary purpose of this research study was to determine if AES is more efficient than human raters. By comparing AES with human raters to score constructed-response items on nursing examinations, information about the following research questions was obtained.

Research question 1: Is AES as effective as human raters for scoring constructed-response items in nursing education in terms of accuracy and reliability? When scoring constructed-response items, the gold standard is considered to be human raters (Latifi, 2016; Shermis, 2014). Research on AES typically includes comparisons with human raters to see how closely AES meets the gold standard of human raters. In this study, the effectiveness of AES was compared with human raters and evaluated in terms of agreement measures and reliability coefficients to determine how closely AES and human raters agreed on the scores.

HR1 and AES. In the initial dataset for development of the AES model, the comparisons between HR1 and AES met the standard of ≥ 0.70 for the reliability coefficients for Quadratic Weighted Kappa κ_q and Pearson r for items #1, 2, and 3. Reliability coefficients for Item #4 were close to meeting the standard of ≥ 0.70 with results at 0.68 for both measures. Agreement measures for items #2, 3, and 4 were high with agreement measures for items #3, and

4 being 1.00 and 0.81 respectively. The overall results for the comparison between AES and HR1 for the initial dataset showed high levels of agreement and reliability, especially for items #2 and 3.

Comparisons between AES and HR1 for the application of the AES model were also high for most items. Reliability coefficients for this comparison were close to or above the standard for items #2, 3, and 4. Agreement measures for items #2, 3, and 4 were over 97 percent for this comparison, which is well above the recommended standard for agreement. The results for the application of the AES model for item #1 showed low results for the comparison with reliability coefficients at ≤ 0.60 for both measures and only 85 percent on agreement measures. Overall, the agreement between AES and HR1 met or were close to meeting the standards for reliability and agreement measures for items #2, 3, and 4 in the initial dataset and the application of the AES model dataset. Although item #1 met the recommended agreement standards in the initial dataset, the results for the application of the AES model were below the recommended standards.

HR2 and AES. Overall the results for the comparisons between HR2 and AES were lower than the comparisons between HR1 and AES. In the initial dataset for the development of the AES model, only item #3 met the recommended standards for reliability and agreement measures. Items #1, 2, and 4 were all below 0.50 for reliability measures and less than 55 percent for agreement measures. These results are likely from the process of using the scores from HR1 to develop the AES model. The scores from HR2 were not used in the development of the AES model from the initial dataset.

In the application of the AES model, the reliability coefficients for comparisons between HR2 and AES were much improved for items #2, 3, and 4 with perfect agreement for item #3 at 1.00. The agreement measures for items #2, 3, and 4 were well above recommended standards at

over 97 percent with item #3 at perfect agreement. The reason for the discrepancy in these results between the initial dataset and the application of the AES model for the comparison of HR2 and AES is likely due to the lower levels of agreement between human raters for the initial dataset when compared with the application of the AES model. This is outlined in the next section.

HR1 and HR2. One of the more interesting findings in this study is the overall lack of agreement between human raters in the initial dataset for the development of the AES model. All of the reliability and agreement measures for items #1, 2, and 4 were well below the recommended standards. Item #3 was the only item that met the recommended standards. Although lack of interrater reliability has been identified as a concern in scoring constructed-response items (Attali, Lewis, & Steier, 2013; Gierl et al, 2014; Kuo, Chen, Yang, & Mok, 2016; Shermis & Burstein, 2013), the level of disagreement between human raters on the initial dataset is concerning. This is an unexpected result for this comparison for the initial dataset and may be attributed to several factors, which will be addressed in the limitations section of this chapter.

In the dataset for the application of the AES model, the agreement and reliability measures were much higher with item #3 again reaching perfect agreement for all measures. Item #2 had over 97 percent agreement measures and slightly below 0.70 for reliability coefficients. Item #4 had over 97 percent agreement measures but the results for reliability coefficients were well below 0.70 which was addressed in chapter 4. Item #1 again had poor results for reliability and agreement, which is possibly due to the challenges with the item and its development. Overall the agreement between humans was much higher for the scores from the 40 students in 2018 when compared with the scores from 2016.

Research question 2: Does AES score constructed-response items more efficiently than human raters? This research question was clearly answered in this study. Overall, HR1 spent 37.33 hours scoring all responses for the study (359 and 40 student responses) and HR2 invested a total of 43 hours (359 and 40 student responses) for scoring. The overall average time for scoring one student response to the four constructed-response items was 5.67 minutes for HR1 and 6.45 minutes for HR2. AES required less than one second to score four constructed-response items for each student. In terms of time, AES is much more efficient than humans for scoring.

This is an expected result in this study. Several research studies have shown that AES can score items much quicker and more efficiently than human raters (Attali, Lew, & Steier, 2012; Gierl et al, 2014; Yang, Liu, & Morell, 2018). One of the important factors to note from this study is the amount of time and data needed to develop an AES model for scoring. It is essential to recognize that larger amounts of data are needed to develop an accurate model for AES (Shermis & Burstein, 2013; Shermis & Morgan, 2016). Shermis and Morgan (2016) suggested that data sets with thousands of entries are more effective for developing an accurate AES model from the training set. Also, data preparation and preprocessing requires time and effort to ensure an accurate AES model is developed from the training set.

When looking at the overall measures of time comparing human raters with AES, it seems apparent that AES is more efficient. Development of the AES model required over 25 hours of human time to preprocess, train, and run the scoring model. Once the model was developed, AES is more efficient but it is essential to consider the amount of time needed up front for AES to work. In consideration of using AES in nursing education assessments, it is critical to ensure development of good constructed-response items that can be used multiple

times over several years. The patient scenario and four constructed-response items used in this research study are reflective of a realistic clinical situation. These items could be used for several years in multiple nursing examinations at various levels of nursing education programs making the overall efficiency of AES much higher than using human raters.

Research question 3: Is AES a potential solution to overcome the challenges of time, cost, and subjectivity in scoring constructed-response items? The results of this study demonstrated that AES is a potential solution for overcoming the challenges in scoring constructed-response items in nursing education assessments. It was clearly demonstrated that AES is much more efficient than human raters to score constructed-response items.

An important finding from this study is the overall difference in time required by HR1 and AES. The amount of time needed by HR1 to score the initial 359 student responses and the 40 subsequent student responses was over 37 hours. If these items were to be included on examinations for 400 more students, a human rater could potentially use almost 38 hours to score these items. This means that the human rater would spend almost a week scoring items rather than engaging with students, teaching, supporting, and exploring teaching and learning strategies. It is also important to note that the four constructed-response items for this study accounted for only 10 marks out of a total of 60 marks on the nursing examinations. This means that human raters invested a large amount of time to score only a small portion of the learning assessments for the nursing students. The findings from this study demonstrate that AES can accomplish this task much more efficiently than human raters, which would allow the educators to spend more time with students to support learning. Even when considering all the time needed for the development of the AES model, it was still more efficient to use AES than human raters to score the constructed-response items. This also addresses the challenge of cost for

scoring constructed-response items. By reducing the amount of time needed to score the items, costs are reduced. A common practice in university education programs is to hire markers. Teaching assistants, graduate students, and other faculty members are hired to score large numbers of learning assessments. Utilizing AES for scoring these assessments would overall decrease the costs associated with scoring these.

More important than time and costs is the effectiveness of AES in reducing subjectivity and increasing reliability in scoring. Some studies have identified subjectivity and lack of reliability between raters as a major concern for university students and faculty (Andrich & Marais, 2018; Minnich, Kirkpatrick, Goodman, Whittaker, Chapple, Schoening, & Khanna, 2018; Saville, 2012). Students report that lack of agreement between faculty members when scoring assessments is a significant concern and major stressor for their overall education (Minnich et al, 2018; Saville, 2012). Students report feelings of frustration when their score is more dependent on the person who scored their work rather than the work itself (Minnich et al, 2018). The findings from this study clearly outline the challenges of interrater reliability when comparing human raters. Overall, AES compared with human raters had higher agreement and reliability measures than human raters compared with each other. Also, AES is objective in scoring. AES does not fatigue or get influenced by other factors, such as emotion or physical wellness. Human raters can be influenced by emotions and how they are feeling at the time. Also, human raters can be affected by what they read in earlier responses. This means that as the human raters read the responses, their scoring continues to be influenced by every entry they read. Another important factor in subjectivity of human raters is giving students marks even when the responses are incorrect. Andrich and Marais (2018) found that human raters often give marks just for effort meaning that even if the student writes a response that is incorrect, humans

often give partial marks to reward the effort by the student. AES scores each item independently based upon the algorithms developed and is not influenced by other factors.

Another significant consideration for using AES in assessing learning in nursing education is the opportunity for timely feedback. Often learners wait days to weeks for feedback on written assessments which then impacts how students apply this feedback. Consider the clinical scenario used in this study. If students received immediate feedback on their performance on the items, they could apply this to actual clinical practice immediately. The students who did well would know that they are on the right track whereas the students who did not do as well would know they need to review this content before going to clinical practice with real patients.

In summary, the findings from this research study suggest that AES is a compelling solution for overcoming the challenges of scoring constructed-response items and may be influential in increasing the use of essay-type items in nursing education assessments.

Second Two: Limitations of this Research Study

There were important limitations in this research study. First, the datasets used for training and development of the AES model and application of the AES model were small. As outlined in many studies on AES (Shermis & Burstein, 2013; Wang & Brown, 2007; Williamson, Xi, & Breyer, 2012), it is essential to use large amounts of data to train and develop an accurate AES model. Although there is no single answer to how much data is needed for the training set, it is important to consider that the more complex the responses, the more data is needed. In order to develop an accurate AES model for these types of constructed-response items, the dataset should be in the thousands of responses (Shermis & Burstein, 2013; Shermis & Morgan, 2016) which would provide a large training set. Shermis and Burstein (2013) identified

the benefit of developing a substantive, comprehensive scoring rubric to develop a corpus of text for the AES model to help overcome smaller datasets. Overall, a larger dataset would have benefitted the training and development of the AES model by giving the program more information to work with to learn the scoring model. In this study, the small datasets impacted the development of the AES model by giving the AES program less information to work with to learn the scoring patterns. The researcher attempted to overcome this by using the information in the scoring rubric to enhance the corpus of text for teaching AES, but overall, a larger dataset would have been valuable. It is important to note that even with the smaller datasets, the results demonstrated the effectiveness of using AES to score constructed-response items.

Another limitation of this study was the scoring rubric. The importance of a well developed, comprehensive scoring rubric is outlined in several studies (Minnich et al, 2018; Kan & Bulut, 2014; Reising, Carr, Tieman, Feather, & Ozdogan, 2015). It is challenging for an educator to think of all possibilities of responses from students. The scoring rubric used in this study was reviewed and tested by several nursing academics and graduate students. Since the items were included on an undergraduate nursing examination, it likely would be more beneficial to have the rubric reviewed by similar level nursing students rather than experienced clinicians (Baldwin, Fowles, & Livingston, 2005; Johnson, Schwartz, Lineberry, Rehman, Soo Park, 2018). This may have resulted in a more comprehensive, broader scoring rubric that could have helped overcome the small size of datasets. The student responses included several points that were not outlined in the scoring rubric, which meant AES had to continue to gather more information for the scoring model. This may have limited the development of the optimal scoring model. Overall, the rubric is an essential aspect of scoring constructed-response items for all

assessments but due to the small datasets used in this study, the impact of the scoring rubric was increased.

Another limitation of this study was the lack of agreement between the human raters. As discussed previously in this chapter, the human raters demonstrated a lower level of agreement and reliability on the initial dataset than expected. This is a limitation because the comparisons and analyses are challenging to understand due to the large differences in scores from human raters. Many factors may have contributed to this. HR1 was involved in the development of the platform for the administration of the examination and was familiar with how the process worked. Although, the process of scoring was duplicated for both human raters, HR2 was less familiar with the scoring platform for the initial dataset. This may account for the low levels of agreement and reliability in the initial dataset when compared with the higher levels of agreement and reliability between human raters in the application of AES model dataset. Both human raters were familiar with the scoring process and platform for the second dataset. This speaks to the overall preparedness of the raters. Rater training and preparedness is essential to improve reliability and agreement between raters for scoring constructed-response items (Baldwin, Fowles, & Livingston, 2005; Johnson et al, 2018; Kan & Bulut, 2014; Minnich et al, 2018). Increasing the preparation for both human raters on the scoring rubric, computer-based scoring process, and items may have increased the agreement between human raters. This may have resulted in higher agreement and reliability measures for all comparisons. Another factor affecting the lack of agreement between humans may be the motivation of the human raters. For both datasets, HR1 was scoring the responses and recording actual scores for the students' grades whereas HR2 was scoring all the responses strictly for the purposes of this research study with no impact on the students' scores. This may have affected the scores by HR1 and HR2. Finally,

expertise and experience of the human raters may have affected the level of agreement. Johnson et al. (2018), Kan and Bulut (2014), and Minnich et al (2018) noted that factors such as expertise and experience affect how humans scored student responses. These factors also affect how humans interpret and comprehend scoring rubrics (Kan & Bulut, 2014). Both human raters in this research study have over 10 years of teaching experience with nursing students, however, HR1 has more experience in scoring constructed-response items than HR2 which may have impacted the results. These factors support the rationale for increasing rater training and developing clear and objective scoring rubrics.

Item development may have also been a limitation of this research study. Item #1 was identified as challenging to score even with the rubric. The overlap of correct responses and several versions of a correct response made item #1 an ineffective item requiring further development. This limitation may have been identified earlier on in this project if undergraduate nursing students were included in the development of the scoring rubric. It may have been evident that there were overlapping correct responses and multiple correct versions of responses if more students had been involved in the item development and testing. In high stakes examinations, items are reviewed, tested, and run through simulations to ensure high quality item development. Of course, most education programs do not have enough resources to commit to those levels of item development but some simple, less costly steps can be included to facilitate development of higher quality items; such as, including questions on exams as non-graded items to gather information about the quality of the item. Also, practice examinations can be given to students with items that can be used on future assessments to identify issues in the practice items. Another simple solution is giving practice items to colleagues and content experts to help improve item quality. Currently, major publishing companies have substantial exam banks

available for educators and learners to use. Accessing some of these items or modeling item development from these exam banks may be helpful. Overall, item development has a major impact on the quality of the examination and committing resources and time is essential to ensure accurate assessment of learning (Johnson et al., 2018; Kuo, Chen, Yang, & Mok, 2016; Yang, Liu, & Morell, 2018). Content experts working with item development experts is the best way to achieve high quality items.

Section Three: Directions for Future Research

There are at least four key directions for future research into using AES to assess higher-level thinking skills in nursing education. This research study explores the effectiveness of AES in scoring constructed-response items on nursing examinations and is a beginning step for future research.

One of the key directions for research is continuing to use AES to score the constructed-response items used in this study. Before continuing in this direction, it is essential to redesign item #1 and use the information gathered in this project to guide development of the item and scoring rubric. Once item #1 is redesigned and reviewed, the four constructed-response items can be included on future nursing examinations or assessments and scored by AES and human raters. It would be valuable to continue analyzing the comparisons of the scores between human raters and AES and replicating this study with larger datasets. The items in this research study can be shared with other nursing education programs to increase the size of the dataset. This would be beneficial for developing an optimal AES model for these items.

A second key direction for future research is developing more constructed-response items and including these on nursing examinations and assessments to begin the process of creating a bank of items that can be scored by AES. The overall goal is not to completely replace human

scoring, but to augment human scoring with AES. For example, items can be scored by humans and checked with AES. Any large discrepancies in scores would indicate the need to look at the response and determine the most accurate score. Another example would be to score all the responses with AES and then randomly score 50 percent of the responses by a human. If all the scores fall within acceptable agreement parameters, the scores by AES stand. If the scores do not fall within acceptable agreement parameters, all the responses would be scored by a human. Each of these examples would be valuable research projects. The constructed-response items can be included in any learning assessment for nursing students, not just examinations. For example, preparation for lab, simulation experiences, and clinical practice could all incorporate AES to score items and provide timely feedback to the students.

A third key direction for research in AES in nursing education is the use of AES to score practice items for preparation materials for licensure examinations. Currently, the licensure examinations for nursing practice are comprised of only selected-response items, however, this is changing to include constructed-response items. This change has recently been announced as a response to the demand for increased assessment of clinical judgement on the nursing licensure examinations. Preparation materials and programs for licensure examinations are offered by many major publishing companies and include tens of thousands of practice items for the students to complete. Most of these preparation programs are digital (electronic) resources. With the proposed inclusion of constructed-response items on the licensure examinations, it would be interesting to incorporate AES into the preparation programs to give the students feedback and scores on their responses. Since the current licensure examinations are used across North America, the datasets for practice constructed-response items would be very large. This would be a valuable research project to explore the usefulness of AES in scoring preparation

materials.

Finally, a fourth key direction would be exploring the students' thoughts and perspectives on the use of AES to score constructed-response items on nursing examinations. Attitude scales, surveys, and open-ended questions could be given to the students to explore their attitudes and thoughts on the use of AES to score their exams. The current generation of nursing students is more trusting of technology when compared with students from years ago. The use of technology in nursing education, nursing practice, and patient care has evolved dramatically and students experience these technological advances daily. A research study on the perceptions of students on the use of AES to score constructed-response items would be informative and helpful to guide nursing education. It would also be helpful to gather information from the students about the inclusion of constructed-response items on nursing examinations and the potential for timely feedback when AES is used.

Conclusions

In the 21st century, we are benefitting from many technological advances. Driverless cars, refrigerators that notify us with a text message on our phone that our milk is low or expiring soon, groceries that are automatically delivered to our doors, and printers that routinely order ink when it is detected that the ink levels are low. Consider being a new university student sitting in a class with 299 other students and being informed that your entire grade will be determined solely on your performance on selected-response item examinations. For some students, this may be great news since the possibility of guessing correctly is always available. For many, this single type of assessment means that much of their learning may not be assessed. Or, consider that you are in the same class of 300 students and you are informed that there will be 25 different humans scoring the learning assessments for your class and you hope that you get the easy

marker. Students may wonder how their refrigerator at home can detect the expiry date on a carton of eggs but important things, like how their grades are determined, are left to the same processes that have been used for decades. Fortunately, AES can change this and the continued development and research into the use of this technology is essential.

There is much support in the literature for the inclusion of both selected-response and constructed-response items in educational assessments and it is well recognized that higher-level thinking skills are more accurately assessed with constructed-response items (Johnson et al, 2018; Kuo, Chen, Yang, & Mok, 2016; Minnich et al, 2018; Tankersley, 2007; Yang, Liu, & Morell, 2018). Higher-level thinking skills in nursing education, such as, critical thinking, clinical reasoning, and clinical judgement, require accurate assessments to ensure effective learning and evaluation which impact safe patient care. The upcoming changes in nursing licensure examinations are based upon evidence of the importance of including constructed-response items in learning assessments. Like many educational programs, nursing class sizes have been increasing which is due mostly to the cost savings associated with larger class sizes. The challenges with scoring constructed-response items are well documented in the literature and AES is a solution that can be used to overcome these challenges.

The main purpose of this research study was to explore the use of AES to assess higher-level thinking skills in nursing education. The results from this study demonstrate that AES is effective and efficient for scoring constructed-response items in nursing educational assessments and that more research is needed in this area. Replication of this study, implementation of new studies, as well as exploring student perspectives on AES are some of the suggestions for future research with AES in nursing education.

References

- Aiken, L., Clark, S., Cheung, R., Sloane, D., & Silber, J. (2003). Educational levels of hospital nurses and surgical patient mortality. *Journal of American Medical Association*, 290 (12), 1617-1623.
- Albon, C. (2018). *Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning*. O'Reilly Media, INC.
- Alfaro-LeFevre, R. (2017). *Critical thinking, clinical reasoning, and clinical judgment: A practical approach (6th ed.)*. Philadelphia, PA: Elsevier.
- Alinier, G. (2010). Developing high-fidelity health care simulation scenarios: A guide for educators and professionals. *Simulation & Gaming*, 42 (1), 9-26.
- Almond, R. G. (2014). Using automated essay scores as an anchor when equating constructed-response writing tests. *International Journal of Testing*, 14(1), 73-91. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1025252&site=eds-live&scope=site;http://dx.doi.org/10.1080/15305058.2013.816309>
- Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, Massachusetts: The MIT Press.
- American Association of Colleges of Nursing (2008). *The essentials of baccalaureate education for professional nursing practice*. Retrieved from www.aacn.nche.edu/education-resources/baccessentials08.pdf
- American Philosophical Association (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Millbrae, CA: The California Academic Press.

- Andrich, D., & Marais, I. (2018). Controlling Bias in Both Constructed Response and Multiple-Choice Items When Analyzed with the Dichotomous Rasch Model. *Journal Of Educational Measurement*, 55(2), 281-307.
- Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1), 125-141. doi:10.1177/0265532212452396
- Attali, Y. & Brustein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4 (3). Retrieved from <http://ejournals.bc.edu/ojs/indes.pho/jtla/>
- Baldwin, D., Fowles, M., & Livingston, S. (2008). Guidelines for constructed-response and other performance assessments. Educational Testing Service. Retrieved from www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf
- Balla, J., Heneghan, C., Thompson, M., & Balla, M. (2012). Clinical decision making in a high-risk primary care environment: a qualitative study in the UK. *BMJ Open*, 2, e000414. doi: 10.1136/bmjopen-2011-000414
- Banerjee, M., Capozzoli, M., McSweeney, L. & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, 27 (1), 3-23.
- Barrett, A. K. & Stephens, K. K. (2017). Making electronic health records (EHRs) work: Informal talk and workarounds in healthcare organizations. *Health Communication*, 32 (8), 1004-1013. doi:10.1080/10410236.2016.1196422
- Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 142-173). New York: Routledge.

- Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355, 2217-2225.
- Bowen, J. L., & Ilgen, J. S. (2014). Now You See It, Now You Don't: What Thinking Aloud Tells Us About Clinical Reasoning. *J Grad Med Educ*, 6(4), 783-785. doi: 10.4300/JGME-D-14-00492.1
- Brookhart, S. M. (1993). Assessing student achievement with term papers and written reports. *Educational Measurement: Issues and Practice*, Spring, 40-47.
- Brookhart, S. M. & Nikito, A. J. (2015). *Educational assessment of students (7th ed.)*. Upper Saddle River, NJ: Pearson Education Inc.
- Burbach, B., Barnason, S., & Thompson, S. A. (2015). Using "Think Aloud" to Capture Clinical Reasoning during Patient Simulation. *International Journal of Nursing Education Scholarship*, 12(1). doi: 10.1515/ijnes-2014-0044
- Burke, H. & Mancuso, L. (2012). Social cognitive theory, metacognition, and simulation learning in nursing education. *Journal of Nursing Education*, 51(10), 543-8.
- Byers, V. (2017). The challenges of leading change in health-care delivery from the front-line. *Journal of Nursing Management*, 25 (6), 449-456.
- Canadian Patient Safety Institute (2008). Patient simulation needs assessment. Retrieved <http://www.patientsafetyinstitute.ca/English/education/simulation/Documents/Patient%20Needs%20Assessment%20-%20May%202008.pdf>.
- Candler, C. (2007). *Effective use of educational technology in medical education*. Paper presented at the Colloquium on educational technology: recommendations and guidelines for medical educators. Washington, DC: AAMC Institute for Improving Medical Education.

- Cappelletti, A., Engel, J. K., & Prentice, D. (2014). Systematic review of clinical judgement and reasoning in nursing. *Journal of Nursing Education, 53* (8), 453-458.
- Carter, A. G., Creedy, D. K., & Sidebotham, M. (2016). Efficacy of teaching methods used to develop critical thinking in nursing and midwifery undergraduate students: A systematic review of the literature. *Nurse Education Today, 40*, 209-218.
- Causser, J., Barach, P., & Williams, A. M. (2014). Expertise in medicine: using the expert performance approach to improve simulation training. *Medical Education, 48*(2), 115-123. doi: 10.1111/medu.12306
- Cazzell, M. & Anderson, M. (2016). The impact of critical thinking on clinical judgement during simulation with senior nursing students. *Nursing Education Perspectives, 37* (2), 83-90.
- Chang, R. W., Bordage, G., & Connell, K. J. (1998). The importance of early problem representation during case presentations. *Academic Medicine, 73*(10 Suppl), S109-111.
- Clynes, M. P. & Raftery, S. E. C. (2008). Feedback: An essential element of student learning in clinical practice. *Nurse Education in Practice, 8*, 405-11.
- Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(1), 100-108. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ995507&site=eds-live&scope=site;http://dx.doi.org/10.1016/j.asw.2012.11.001>
- Conrad, M. A., Guhde, J., Brown, D., Chronister, C., & Ross-Alaolmolki, K. (2011). Transformational leadership: Instituting a nursing simulation program. *Clinical Simulation in Nursing, 7*, 189-195.

- Corrigan, R. & Hardham, G. (2011). Use of technology to enhance student self evaluation and the value of feedback on teaching. *International Journal of Therapy & Rehabilitation, 18*(10), 579-590.
- Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022-1028. doi: 10.1097/ACM.0b013e3181ace703
- Curtin, L. B., Finn, L. A., Czosnowski, Q. A., Whitman, C. B., & Cawley, M. J. (2011). Computer-based simulation training to improve learning outcomes in mannequin-based simulation exercises. *American Journal of Pharmaceutical Education, 75*(6), 1-113.
Retrieved from <http://search.proquest.com/docview/892735918?>
- Cushing, A., Abbott, S., Lothian, D., Hall, A., & Westwood, O. (2011). Peer feedback as an aid to learning -- what do we want?? feedback. when do we want it?? now! *Medical Teacher, 33*(2), e105-12. doi:10.3109/0142159X.2011.542522
- del Bueno, D. (2005). A crisis in critical thinking. *Nursing Education Perspectives, 26*(5), 278-282.
- DeMars, C. E. (2008). Scoring multiple-choice items: A comparison of IRT and classical polytomous and dichotomous methods. Retrieved from <http://www.jmu.edu/assessment/CED%20NCME%20Paper%2008.pdf>
- Dexter, P., Applegate, M., Backer, J., Claytor, K., Keffer, J., Norton, B., & Ross, B. (1997). A proposed framework for teaching and evaluating critical thinking in nursing. *Journal of Professional Nursing, 13* (3), 160-167.
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1) Retrieved from

<http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ843855&site=eds-live&scope=site>

- Dillon, P. M., Noble, K. A., & Kaplan, W. (2009). Simulation as a means to foster collaborative interdisciplinary education. *Nursing Education Perspectives, 30*(2), 87–90.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19* (2), 143-165.
- Edwards, S. L. (2007). Critical thinking: A two-phase framework. *Nursing Education in Practice, 7*, 303-314.
- Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine, 15*(11), 988-994.
- Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical Education, 39*(1), 98-106. doi: 10.1111/j.1365-2929.2004.01972.x
- Field, A. (2018). *Discovering statistics using IBM® SPSS® Statistics: North American version*. Thousand Oaks, CA: Sage Publications Inc.
- Fleming, A., Cutrer, W., Reimschisel, T., & Gigante, J. (2012). You too can teach clinical reasoning! *Pediatrics, 130*(5), 795-797. doi: 10.1542/peds.2012-2410
- Foltz, P. W. (2016). Advances in automated scoring of writing for performance assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 658-677). Hershey, PA: Information Science Reference, IGI Global.

- Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. New York: Cambridge University Press.
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A., & Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, (10), 950. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.398149659&site=eds-live&scope=site>
- Gierl, M. J. (2014). Lecture notes, Test Theory, Educational Psychology 507, University of Alberta, unpublished.
- Goodman, J. S., Wood, R. E., & Chen, Z. (2011). Feedback specificity, information processing, and transfer of training. *Organizational Behavior and Human Decision Processes*, 115(2), 253-267. Retrieved from <http://search.proquest.com/docview/881465434?>
- Gordon, J. (2000). Congruency in defining critical thinking by nurse educators and non-nurse scholars. *Journal of Nursing Education*, 39 (8), 340-351.
- Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. *Center for Educator Compensation Reform*, Feb (2012), 1-33.
- Greidanus, E, King, S., LoVerso, T., & Ansell, L. D. (2013). Interprofessional learning objectives for health team simulations. *Journal of Nursing Education*, 52 (6), 311-6.
- Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics*, 35(5), 586-602. Retrieved from

[http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ951762&site=eds-live&scope=site;
http://dx.doi.org/10.3102/1076998610375839](http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ951762&site=eds-live&scope=site;http://dx.doi.org/10.3102/1076998610375839)

- Hall, O. L., Soderstrom, T., Ahlqvist, J., & Nilsson, T. (2011). Collaborative learning with screen-based simulation in health care education: An empirical study of collaborative patterns and proficiency development. *Journal of Computer Assisted Learning, 27*(5), 448-461.
- Harmes, J. C., Welsh, J. L., & Winkelman, R. J. (2016). A framework for defining and evaluating technology integration in the instruction of real-world skills. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 137-162). Hershey, PA: Information Science Reference, an imprint of IGI Global.
- Harmon, M. M. & Thompson, C. (2015). Clinical reasoning in pre-licensure nursing students. *Teaching and Learning in Nursing, 10*, 63-70.
- Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measure. *Journal of Applied Behavior Analysis, 10*, 103-116.
- Huckaby, L. (2009). Clinical reasoned judgement and the nursing process. *Nursing Forum, 44*, 72-78.
- Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27* (1), 10-28.
- Johnson, J., Schwartz, A., Lineberry, M., Rehman, F., & Park, Y. S. (2018). Development, administration, and validity evidence of a subspecialty preparatory test toward licensure:

- A pilot study. *BMC Medical Education*, 18doi:http://dx.doi.org/10.1186/s12909-018-1294-z
- Kahneman, D. (2011). *Thinking, fast and slow*: Doubleday Canada.
- Kan, A., & Bulut, O. (2014). Crossed Random-Effect Modeling: Examining the Effects of Teacher Experience and Rubric Use in Performance Assessments. *Eurasian Journal Of Educational Research*, (57), 1-27.
- Kassirer, J. P. (1995). Teaching problem-solving--how are we doing? *New England Journal of Medicine*, 332(22), 1507-1509. doi: 10.1056/NEJM199506013322210
- Kellogg, R.T. & Raulerson III, B.A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, 14 (2), 237-242.
- Koharchik, L. Culleiton, A. L., Caputi, L. & Robb, M. (2015). Fostering clinical reasoning in nursing students. *American Journal of Nursing* 115 (1), 58-61.
- Kuiper, R. & Pesut, D. J. (2004). Promoting cognitive and metacognitive reflective clinical reasoning skills in nursing practice: Self-regulated learning theory. *Journal of Advanced Nursing*, 45 (4), 381-391.
- Kuiper, R., Pesut, D., & Kautz, D. (2009). Promoting the self-regulation of clinical reasoning skills in nursing students. *The Open Nursing Journal*, 3, 76-85.
- Kuo, B. C., Chen, C. H., Yang, C. W., & Mok, M. M. C. (2016) Cognitive diagnostic models for tests with multiple-choice and constructed-response items, *Educational Psychology*, 36:6, 1115-1133, DOI: 10.1080/01443410.2016.1166176
- Lam, C. F., DeRue, D. S., Karam, E. P., & Hollenbeck, J. R. (2011). The impact of feedback frequency on learning and task performance: Challenging the “more is better” assumption. *Organizational Behavior and Human Decision Processes*, 116, 217-228.

- Lapkin, S., Fernandez, R., Levett-Jones, T., & Bellchambers, H. (2010). *The effectiveness of using human patient simulation mannequins in the teaching of clinical reasoning skills to undergraduate nursing students: A systematic review*. Adelaide, Australia. Retrieved from <http://search.proquest.com/docview/750364569?>
- Latifi, S. M. F. (2016). Development and validation of an automated essay scoring framework by integrating deep features of English language. Dissertation. University of Alberta: Edmonton, AB.
- Latifi, S. M. F., Gierl, M. J., & Boulais, A. P. (2013). *Towards the use of an automated scoring framework at the Medical Council of Canada: An exploratory approach*. Medical Council of Canada
- Lindsay, P. L. & Jenkins, S. (2013). Nursing students' clinical judgement regarding rapid response: The influence of a clinical simulation education intervention. *Nursing Forum*, 48 (1), 61-70.
- Mayfield, E. & Rose, C. P. (2013). LightSIDE: Open source machine learning for text. In M.D. Shermis & J.C. Burstein (Eds.) *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124-135). Routledge.
- McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, 15(2), 118-129. doi:10.1016/j.asw.2010.04.002
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23, 35-59. doi:10.1016/j.asw.2014.09.002

- Minnich, M., Kirkpatrick, A. J., Goodman, J. T., Whittaker, A., Chapple, H. S., Schoening, A. M., & Khanna, M. M. (2018). Writing Across the Curriculum: Reliability Testing of a Standardized Rubric. *Journal of Nursing Education, 57*(6), 366-370.
- Motola, I., Devine, L. A., Chung, H. S., Sullivan, J. E., & Issenberg, S. B. (2013). Simulation in healthcare education: A best evidence practical guide. AMEE guide no. 82. *Medical Teacher, 35*(10), E1511. Retrieved <http://search.proquest.com/docview/1437347219?>
- Murphy, J. L. (2004). Using focused reflection and articulation to promote clinical reasoning: An evidence-based teaching strategy. *Nursing Education Perspectives, 25* (5), 226-231.
- Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education, 44*(1), 94-100. doi: 10.1111/j.1365-2923.2009.03507.x
- Oermann, M. H. & Gaberson, K. B. (2017). *Evaluation and testing in nursing education (5th ed.)*. New York: Springer Publishing Company.
- Oestergaard, J., Bjerrum, F., Maagaard, M., Winkel, P., Larsen, C. R., Ringsted, C., Soerensen, J. L. (2012). Instructor feedback versus no instructor feedback on performance in a laparoscopic virtual reality simulator: A randomized educational trial. *BMC Medical Education, 12*, n/a-7. doi:<http://dx.doi.org/10.1186/1472-6920-12-7>
- O'Neill, E. S. & Dluhy, N. M. (1997). A longitudinal framework for fostering critical thinking and diagnostic reasoning. *Journal of Advanced Nursing, 26*, 825-832.
- Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates
- Parekh, A., & Thorpe, T. (2012). How should we teach undergraduates in simulation scenarios? *The Clinical Teacher, 9*(5), 280-284. doi:10.1111/j.1743-498X.2012.00552.x

- Pelaccia, T., Tardif, J., Tribby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Med Educ Online, 16*. doi: 10.3402/meo.v16i0.5890
- Posel, N., McGee, J. B., & Fleiszer, D. M. (2014). Twelve tips to support the development of clinical reasoning skills using virtual patient cases. *Medical Teacher, 1-6*. doi: 10.3109/0142159X.2014.993951
- Python Software Foundation (2018). Python software 3.7.0. Retrieved from <https://www.python.org/downloads/>
- Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25-39. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ995506&site=eds-live&scope=site;http://dx.doi.org/10.1016/j.asw.2012.10.004>
- Raymond-Seniuk, C. & Profetto-McGrath, J. (2011). Can one learn to think critically?--A philosophical exploration. *Open Nursing Journal, 5*, 45-51.
- Reilly, E. D., Stafford, R. E., Williams, K. M., & Brooks Corliss, S. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open & Distance Learning, 15*(5), 84-98. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=99397695&site=eds-live&scope=site>

- Reising, D. L., Carr, D. E., Tieman, S., Feather, R., & Ozdogan, Z. (2015). Psychometric Testing of a Simulation Rubric for Measuring Interprofessional Communication. *Nursing Education Perspective*, (5), 311. doi:10.5480/15-1659
- Rencic, J. (2011). Twelve tips for teaching expertise in clinical reasoning. *Medical Teacher*, 33(11), 887-892. doi: 10.3109/0142159X.2011.558142
- Research Ethics Office (2016). REB 2. Retrieved from <http://www.reo.ualberta.ca/Human-Research-Ethics/Research-Ethics-Boards/REB-2.aspx>
- Rudner, L. R. & Liang, T. (2002). Automated essay scoring using Bayes' Theorem. *Journal of Technology, Learning, and Assessment*, 1 (2). Retrieved from www.jtla.org.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric monograph No. 17*. Richmond, VA: Psychometric Society.
- SAS (2016). Machine learning. What it is & why it matters. *SAS: The power to know*. Retrieved from http://www.sas.com/it_it/insights/analytics/machine-learning.html
- Sangwin, C. & Jones, I. (2017). Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics*, 94(2), 205-222. doi:10.1007/s10649-016-9725-4
- Saville, B. K. (2012). The ethics of grading. In R. E. Landrum, M. A. McCarthy, R. E. Landrum, M. A. McCarthy (Eds.), *Teaching ethically: Challenges and opportunities* (pp. 31-42). Washington, DC, US: American Psychological Association. doi:10.1037/13496-003
- Schlegel, C., Woermann, U., Rethans, J., & van der Vleuten, C. (2012). Validity evidence and reliability of a simulated patient feedback instrument. *BMC Medical Education*, 12, 6. doi:<http://dx.doi.org/10.1186/1472-6920-12-6>

Scriven, M. & Paul, R. (2013). Defining critical thinking. *The Critical Thinking Community*.

Retrieved from <http://www.criticalthinking.org/pages/defining-critical-thinking/766>

Shellenbarger, T. & Robb, M. (2015). Technology-based strategies for promoting clinical reasoning skills in nursing education. *Nurse educator, 40* (2), 79-82.

Sireci, S. G. & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development* (pp. 329-347). Lawrence Erlbaum Associates, Inc.: New Jersey.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Shermis, M. & Barrera, F. (2002). Exit assessments: Evaluation of writing ability through automated essay scoring (ERIC document reproduction service no. ED 464 950).

Shermis, M. D. & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D. & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M. D., & Morgan, J. (2016). Using prizes to facilitate change in educational assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 323-338). New York: Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Simmons, B. (2010). Clinical reasoning: Concept analysis, *Journal of Advanced Nursing, 66*, 1151-1158.

- Stegmann, K., Pilz, F., Siebeck, M., & Fischer, F. (2012). Vicarious learning during simulations: Is it more effective than hands-on training? *Medical Education*, *46*(10), 1001-1008.
doi:10.1111/j.1365-2923.2012.04344.x
- Stemler, S. E. (2004). Automated essay scoring: A human's review. *Psyccritiques*, *49*
doi:10.1037/04098S
- Stemler, S. E. (2004a). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9* (4), 135-144.
- Tan, X., Kim, S., Paek, I., & Xiang, B. (2009). An alternative to the trend scoring method for adjusting scoring shifts in mixed-format test. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA.
- Tankersley, K. (2007). Constructed response: Connecting performance and assessment. Tests that teach. *Association for Supervision and Curriculum Development*, *1*, 1-9.
- Tanner, C. A. (2006). Thinking like a nurse: A research-based model of clinical judgement in nursing. *Journal of Nursing Education*, *45*, 204-211.
- Techopedia (2018). Retrieved from <https://www.techopedia.com/definition/14650/data-preprocessing>
- Turkel, M. C. & Morrison, D. (2016). Describing self-reported assessments of critical thinking among practicing medical-surgical registered nurses. *MEDSURG Nursing*, *25* (4), 244-251.
- Walkow, J. C., & Reilly, E. (2014). Are we ready for robots to grade? *Chronicle of Higher Education*, *61*(3), B41-B42. Retrieved from

<http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=98380568&site=eds-live&scope=site>

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, 6(2) Retrieved from

<http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ838612&site=eds-live&scope=site;>

<http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1100&context=jtla>

Wang, J., & Michelle, S. B. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology & Teacher Education*, 8(4), 310-325. Retrieved from

<http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=36573682&site=eds-live&scope=site>

WEKA (2015). The University of Waikato, Machine Learning Group. Retrieved from

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues & Practice*, 31(1), 2-13.

doi:10.1111/j.1745-3992.2011.00223.x

Witten, I. H., Frank, E. & Hall, M. A. (2013). Data transformations. In I. H. Witten, E. Frank, & M. A. Hall (Eds.) *Data mining, 3rd edition* (pp. 305-350). Burlington, MA: Elsevier.

Yarnold, P. R. (2016). ODA vs. π and κ : Paradoxes of Kappa. *Optimal Data Analysis*, 5, 160-161.

- Yang, J. S., Liu, Y., & Morell, M. (2018). Constructed-Response Items. In B. B. Frey (Ed.) *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 380-383). Thousand Oaks, CA: SAGE Publications.
- Zori, S. & Morrison, B. (2009). Critical thinking in nurse managers. *Nursing Economic\$, 27 (2)*, 75-79, 98.
- Zuriguél Pérez, E. Lluch Canut, M. T., Falco Pegueroles, A. Puig Llobet, M., Moreno Arroyo, C., & Roldan Merino, J. (2014). Critical thinking in nursing: Scoping review of the literature. *International Journal of Nursing Practice, 21*, 820-830.

Appendix ANursing and Health Sciences programs approached to access constructed-response items:

Faculty of Nursing, University of Alberta

Faculty of Pharmacy, University of Alberta

Faculty of Rehabilitation Medicine, University of Alberta

Faculty of Nursing, MacEwan University

Department of Health Sciences, NorQuest College

Department of Nursing, Athabasca University

Department of Physiology, Athabasca University

Faculty of Nursing, Windsor University, Windsor, ON

Faculty of Nursing, University of Saskatchewan

Faculty of Nursing, Camosun College, Victoria, BC

Faculty of Nursing, Lander University, Greenwood, South Carolina, USA

Faculty of Nursing, Eastern Kentucky University, Kentucky, USA

Faculty of Nursing, Gwynedd Mercy University, USA

Faculty of Nursing, University of Massachusetts, USA

Faculty of Nursing, Indiana University Northwest, USA

Faculty of Nursing, Widener University, USA

Faculty of Nursing, Gadsden State Community College, USA

Elsevier Canada

Wolters Kluwer Health/Lippincott, Williams, & Wilkins

All the above contacts informed me that they no longer use constructed-response items because they are too time consuming and costly to mark. The only examples of constructed-response items that were available to me were from class sizes of less than 20. Several of the programs reported that they have not used short-answer questions for over 10 years due to the cost and time associated with scoring the items.

Appendix C

Scoring Rubric Development—Responses from Colleagues

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data that would be helpful to collect (4 marks).

- Recent history of head trauma
- Medication usage (e.g. anti-coagulants although she is young to be on these there may be comorbidities)
- Family history of TIA, CVA, previous history of same
- Precipitating factors (i.e. working, environmental hazards, stress etc.)
- History of vision status, does she wear glasses or contact lenses?

2) Identify which physical assessments you should perform (3 marks).

- Comprehensive set of neural vitals to see if there are any motor/sensory deficits as a result of her “headache” which could indicate unilateral weakness
- Vital signs, specifically BP to see if she has a hypertensive crisis that could be contributing to the pain

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

- Stat CT head to r/o intracranial bleed and/or ischemia
- Baseline ECG
- Routine bloodwork including PTT/INR

4) Identify possible interventions to help Mrs. S (1 mark).

- Until the CT is performed, and even afterwards, attempts should be made to minimize ICP such as providing a calm, quiet, dark environment that decreases stimuli
- Tornado and Metoclopramide have proven effective for the treatment of the migraines
- Once an intracranial bleed has been ruled out consider giving her some ASA as many OTC combinations for headaches/migraines include acetaminophen and acetylsalicylic acid

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data that would be helpful to collect (4 marks).

- Type, location and intensity of pain
- Nausea, vomiting, light-headedness.
- Any fall or trauma to head – ? SA haemorrhage
- Signs of infection - ? Meningitis
- History of migraine/headaches
- Medication history
- Cognitive, sensory and motor function changes
- Menstrual history
- Is the headache associated with any food or drink?

2) Identify which physical assessments you should perform (3 marks).

- VS monitoring –? Increased temperature or BP
- Sensory exam, reflexes
- Assess of mental status
- Assess balance and muscle strength

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

- Blood test
- CT scan
- MRI

4) Identify possible interventions to help Mrs. S (1 mark).

- Rest
- Analgesics – Acetaminophen, NSAIDs, opioids
- Rule out the cause of headache and treat accordingly.

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

- 1) List the health history data that would be helpful to collect (4 marks).
 - Past medical history such as hypertension, diabetes, HIV
 - Recent injuries or head trauma
 - History of similar headache
 - Family history
 - Medications, including supplements and OTCs
 - Recent traveling
 - Smoker, alcohol or recreational drugs use

- 2) Identify which physical assessments you should perform (3 marks).
 - Vital signs and orientation
 - describe the pain, PQRT, unilateral or bilateral, location, severity, onset and character
 - neuro examination, pupils reactivity, CN strength, light touch sensation, deep tendon reflexes, finger to toe
 - any visual disturbance, motor deficits, paresthesia or language difficulties
 - any associated symptoms, such as recent fever, neck pain, numbness, weakness, nausea or vomiting, photophobia

- 3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).
 - Blood work including electrolytes, WBC, CBCD, glucose, creatinine, GFR, HCG
 - possible CT head
 - can consider MRI if needed
 - lumbar puncture if indicated

- 4) Identify possible interventions to help Mrs. S (1 mark).
 - Interventions will be based on the history, differential diagnosis, lab tests and imaging
 - Fluid boluses if not contraindicated
 - Dark, quiet room if it is migraine headache
 - NSAID or Tylenol if no contraindication
 - Narcotic if analgesic ineffective
 - Antiemetic if needed
 - Use migranal, imitrex, maxeran if needed

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

- 1) List the health history data that would be helpful to collect (4 marks).

Current Hx: query recent sinus/ear infections, trauma/accidents, surgeries

PMHx: query HTN, dyslipidemia, diabetes, cardiac Hx, Ca, IV drug use, migraine headaches

Meds: query anticoagulation therapies

- 2) Identify which physical assessments you should perform (3 marks).

Full neuroassessment: cranial nerves, LOC changes, head and neck, speech, arm drift

Neuromuscular: bilateral assessment of movement and strength and sensation

Cardiovascular: HR, BP, carotid auscultation

- 3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

CT head, carotid U/S, cerebral angiogram, head MRI, Echocardiogram, CBC/lytes, INR/PTT

- 4) Identify possible interventions to help Mrs. S (1 mark).

CT head to determine antithrombolytic appropriateness, antihypertensives

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

- 1) List the health history data that would be helpful to collect (4 marks).

-PQRST for the pain. Other symptoms involved in the headache: Photophobia, phonophobia, nausea, vomiting, weakness, vision changes. What was she doing when at the onset of headache- exercise, reading? Was she feeling well before the headache?

Any aura?

Previous headache history including: previous migraines, tension headaches. Does she have any triggers for headaches that she knows of?

What medications has she tried, if any, for the headaches? If she had this before what did she do to treat it? Did she respond to medication?

General health history: other comorbidities, medication, family history, social history.

- 2) Identify which physical assessments you should perform (3 marks).

Quick neurology exam: cranial nerves including vision assessment (visual fields, vision, fundi), motor, sensory. Neck mobility.

- 3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

Basic labs: CBC-D, electrolytes. If this is a sudden change an MRI may be necessary.

- 4) Identify possible interventions to help Mrs. S (1 mark).

Pain medication for headache- standard IV treatment for migraine and rehydration (if dehydrated).

Education about migraines.

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

Answered by Nurin Dhanji

1) List the health history data that would be helpful to collect (4 marks).

Previous headaches

Medication history

Comorbidities/ other health conditions

Factors leading up to the headache

2) Identify which physical assessments you should perform (3 marks).

Full head to toe

Neurological assessment

Pain assesmenet

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

CT scan HEAD

Full panel blood work

MRI

4) Identify possible interventions to help Mrs. S (1 mark).

Hot/ cold pack

Pain medications

Resting position/ deep breathing technique and relaxation

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data that would be helpful to collect (4 marks).

-Has this happened to you before?

-What medical conditions do you have?

-Family history of headaches of this severity?

-Associated symptoms aside from headache? Recent changes to lifestyle, exposure to certain agents, illness?

2) Identify which physical assessments you should perform (3 marks).

-Neurological

-Cardiovascular Assessment

-Pain assessment

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

-Electrolytes, CBC,

4) Identify possible interventions to help Mrs. S (1 mark).

-Analgesic

-Hydration

-Ongoing monitoring (depending on severity)

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data that would be helpful to collect (4 marks).

Has she taken any medication for it, what and when?

Is she taking any other medications/prescriptions?

Is this the first time for a headache like this?

What was she doing when it started?

Any trauma recently?

Associated symptoms?

Medication allergies?

Other health history or diagnoses? Family history

2) Identify which physical assessments you should perform (3 marks).

Pain assessment (PQRST)

Neuro assessment – GCS, LOC, strength, motor, sensation, pupils

Vitals –HR, BP, RR, O2Sat

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

Blood work – CBCD, Lytes, glucose

CT head

4) Identify possible interventions to help Mrs. S (1 mark).

Symptom management-non sedating analgesic

Determine source of headache

Mrs. S is a 36 year old female who presents at the clinic reporting a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data that would be helpful to collect (4 marks).

Hypertension, neurological disorders, cancers, hx of stroke, TIA, diabetes, family histories for cancers, and neurological disorders, history of mental disorders, current medications, any use of OTC and herbs.

Any factors to make pain worse or relieve...

Any recent injuries, falls, or accidents. Any recent appetite changes, sleep pattern changes...

2) Identify which physical assessments you should perform (3 marks).

Situation description about what happened. Glasgow coma scale, Blood sugars, mini mental scales, visual field or visual acuity check if there is a concern. Inspect, palpate head. Vital signs BP, HR, pain assessment (PQRST). Related cranial nerves check if relevant. Medical history and current medication

3) List all possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

Vital signs, ECG, CBC with differential, urine dips, CT, MRI if there is a concern for head injuries or stroke.

4) Identify possible interventions to help Mrs. S (1 mark).

Analgesic and treat with underline causes.

Appendix D

Scoring Rubric for Constructed-response Item for Automated Essay Scoring Analysis

Mrs. S is a 36 year old female patient on your unit who informs you that she has a “really bad headache”. She reports that her headache started about 1 hour ago and that her head “feels like it is going to explode”.

1) List the health history data (12 items) that would be helpful to collect (3 marks).

- medications
- allergies
- recent travel
- medical conditions/disorders or diabetes
- history of hospitalizations
- previous surgeries/childbirth
- current health status (pregnant)
- history of headache
- family history
- PQRST or OLDCARTS or symptom analysis
- onset or timing
- duration
- quality or nature
- severity or scale of 0-10
- effect on daily activity or significance to client
- alleviating factors
- aggravating factors
- other or associated symptoms
- client perspective
- environmental factors
- location
- pain assessment
- recent injuries or recent falls or recent accidents
- sleep changes

4 items = 1 mark; Any 12 of the above list = 3 marks

2) Identify which physical assessments (8 items) you should perform (2 marks).

- vital signs
- level of consciousness
- FAST
- Stroke assessment
- neurovital signs or Glasgow coma scale
- arm drifting

- facial symmetry
- gait
- Cincinnati stroke test
- talk, wave, smile
- handgrip
- leg movement
- Pupillary response
- talk
- cranial nerve
- swallow
- smile
- frown or facial movement
- visual fields
- visual acuity

4 items = 1 mark; Any 8 of the above list = 2 marks

3) List possible diagnostic tests that should be prescribed for Mrs. S (1 mark).

- blood work or CBC or lytes or differential
- CT scan
- MRI
- ECG
- urinalysis for glucose

2 items = 1 mark; Any 2 of the above list = 1 mark

4) Identify possible nursing interventions (8 items) to help Mrs. S (4 marks).

- positioning
- semi fowlers
- keep NPO or nothing by mouth
- start IV
- oxygen or O₂
- analgesics or pain management
- side rails
- safety
- call family or significant others for support
- reassurance or calm or reduce anxiety
- prepare for TPA
- aspirin or anticoagulant

2 items = 1 mark; Any 8 of the above list = 4 marks

Appendix E

Ethics Approval Application and Approval

Ethics Application has been Approved

ID: [Pro00071611](#)
Title: Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education
Study Investigator: [Tracey Stephen](#)

This is to inform you that the above study has been approved.

Description: Click on the link(s) above to navigate to the HERO workspace.

Please do not reply to this message. This is a system-generated email that cannot receive replies.

University of Alberta
Edmonton Alberta
Canada T6G 2E1

© 2008 University of Alberta
[Contact Us](#) | [Privacy Policy](#) | [City of Edmonton](#)

Appendix FPermission to Access Student Responses from the Faculty of Nursing, U of A

Tracey Stephen
tcs@ualberta.ca

March 3, 2017

Dr. Joanne Profetto-McGrath, Professor and Vice Dean
Faculty of Nursing, University of Alberta

Dear Dr. Joanne Profetto-McGrath

I am currently completing my doctoral studies in Educational Psychology with a specialization in Measurement, Cognition, and Evaluation in the Faculty of Education at the University of Alberta. My research area is focused on using technology to score responses to short answer exam questions. I am currently working on a research study titled, “Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education”.

The research project I am working on requires analysis of multiple responses to short answer questions. These responses are input to a computer software program that automatically scores the responses. It is based on the same notion of computers scoring multiple-choice questions (scantron) only this project focuses on computers scoring short answer type questions.

As you are aware, it is challenging to assess higher-level thinking skills in nursing students using *only* multiple-choice exams. In April 2016, three teaching teams at the Faculty of Nursing incorporated short essay questions into the exams that were administered to students. The students were given a patient scenario then asked 4 questions about the scenario. The incorporation of these questions was done in response to looking at ways to assess critical thinking in addition to multiple-choice exams. These exams were administered through *eclass* and all the responses were recorded and stored.

I have received ethics approval from the University of Alberta Research Ethics Board to run this study. I am requesting permission to access the scored student responses to the questions from April 2016. These exams are already scored and calculated into the students’ grades therefore this study has no impact on student scores related to the use of these data—many of these students graduated in December 2016. Also, all student identifying data will be removed from the responses so that only the actual responses are accessed. There will be no possible way to access any student identifying information.

I have attached a copy of my proposal for you to review. Thank you for your consideration of this important study affecting future nursing education and the use of technology for assessment of learning.

Regards,
Tracey Stephen

Appendix F cont'dPermission to Access Student Responses—Email to Vice Dean, Faculty of Nursing, U of A

Dear Dr. Joanne Profetto-McGrath

I am currently completing my doctoral studies in Educational Psychology with a specialization in Measurement, Cognition, and Evaluation in the Faculty of Education at the University of Alberta and have successfully passed my candidacy. My research area is focused on using technology to score responses to short answer exam questions. I am currently working on a research study titled, "Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education".

The research project I am working on requires analysis of multiple responses to short answer questions. These responses are input to a computer software program that automatically scores the responses. It is based on the same notion of computers scoring multiple-choice questions only this project focuses on computers scoring short answer type questions.

As you are aware, it is challenging to assess higher-level thinking skills in nursing students using *only* multiple-choice exams. In April 2016, three teaching teams at the Faculty of Nursing incorporated short essay type questions into the exams that were administered to students. The students were given a patient scenario then asked 4 questions about the scenario. The incorporation of these questions was done in response to looking at ways to assess critical thinking in addition to multiple-choice exams and also in response to direction from the leadership team. These exams were administered through eclass and all the responses were recorded and stored.

I have received ethics approval from the University of Alberta Research Ethics Board to run this study. The permission approval id number is Pro00071611 and is included in the attachment below. I am requesting permission to access the scored student responses to the questions from April 2016. These exams are already scored and calculated into the students' grades therefore this study has no impact on student scores related to the use of these data--many of these students graduated in December 2016. Also, all student identifying data will be removed from the responses so that only the actual responses are accessed. There will be no possible way to access any student identifying information.

I've attached a copy of my proposal for you to review. I would be happy to meet with you to discuss this study and ensure that all considerations are attended to. Thank you for your consideration of this important study affecting future nursing education and the use of technology for assessment of learning.

Regards,

Tracey Stephen

Appendix F cont'd

Permission to Access Student Responses from Faculty of Nursing, U of A

Hi Tracey,
No worries, it happens. Thanks for updating.

Thank you also for the answer to my question re students' identifying information. You have met all the conditions so please proceed.

All the best,

Joanne

Joanne Profetto-McGrath PhD, RN
Professor & Vice Dean

Appendix G**Code for Data Preprocessing for Python Developed and Used with Permission by Shin (2018)**

```
#!/usr/bin/env python3

# -*- coding: utf-8 -*-

"""

Created on Fri Mar 23 11:22:53 2018

@author: jinnie

"""

from __future__ import print_function

from sklearn.model_selection import train_test_split

import numpy as np

import pandas as pd

import nltk.data

import re

from nltk.corpus import stopwords

from sklearn.metrics import cohen_kappa_score

from nltk.stem import WordNetLemmatizer

from nltk.stem import SnowballStemmer

lemmatizer = WordNetLemmatizer()

#from sklearn import svm

#from sklearn.model_selection import GridSearchCV

#from sklearn.model_selection import train_test_split

#%%%

training_df = pd.read_excel('answer_item3.xlsx').dropna()

resolved_score = training_df['HR #1']

#essay_ids = training_df[training_df['essay_set'] == 1]['essay_id']

responses = training_df['response']

resolved_score = np.reshape(resolved_score, (len(resolved_score), 1))
```

```

#%%%
def clean_str(text, remove_stopwords=False, stem_words=False):

    # Clean the text, with the option to remove stopwords and to stem words.

    # Convert words to lower case and split them
    text = text.lower().split()

    # Optionally, remove stop words
    if remove_stopwords:
        stops = set(stopwords.words("english"))
        text = [w for w in text if not w in stops]
        text = " ".join(text)

    # Clean the text
    text = re.sub(r"^[A-Za-z0-9(),!?\`]", " ", str(text))
    text = re.sub(r"\'s", " \'s", str(text))
    text = re.sub(r"\'ve", " \'ve", str(text))
    text = re.sub(r"n\'t", " n\'t", str(text))
    text = re.sub(r"\'re", " \'re", str(text))
    text = re.sub(r"\'d", " \'d", str(text))
    text = re.sub(r"\'ll", " \'ll", str(text))
    text = re.sub(r",", " ,", str(text))
    text = re.sub(r"!", " !", str(text))
    text = re.sub(r"(", " (", str(text))
    text = re.sub(r")", " )", str(text))
    text = re.sub(r"?", " ?", str(text))
    text = re.sub(r"s{2,}", " ", str(text))
    text = re.sub(r"n", " ", str(text))
    text = re.sub(r"0-9", " ", str(text))
    #text = re.sub(r"\\(", " ( " str(text))
    #text = re.sub(r"\\)", " ) " str(text))

```

```
# Optionally, shorten words to their stems
if stem_words:
    text = text.split()

    stemmer = SnowballStemmer('english')

    stemmed_words = [stemmer.stem(word) for word in text]

    text = " ".join(stemmed_words)

# Return a list of words
return(text)

#%%

response= responses.tolist()
clean=[]
for i in range(len(response)):
    l= clean_str(responses[i])
    clean.append(l)

#%%

wc = []
sc = []
nc = []
vc = []
advc = []
adjc = []
lwc = []
cc = []
pc = []
ld = []
qc = []
```

```
kc1=[]
```

```
kc2=[]
```

```
kc3=[]
```

```
kc4=[]
```

```
def wordCount(text):
```

```
    tokens = nltk.word_tokenize(text)
```

```
    return (len(tokens))
```

```
def sentCount(text):
```

```
    tokens = nltk.sent_tokenize(text)
```

```
    return (len(tokens))
```

```
def nounCount(text):
```

```
    cnt_nn=0
```

```
    tokens = nltk.word_tokenize(text)
```

```
    postags = nltk.pos_tag(tokens)
```

```
    for j in range(len(postags)):
```

```
        if postags[j][1] == "NN" or postags[j][1] == "NNS" or postags[j][1] == "NNP" or postags[j][1] == "NNPS":
```

```
            cnt_nn+=1
```

```
    return (cnt_nn)
```

```
def verbCount(text):
```

```
    cnt_vrb=0
```

```
    tokens = nltk.word_tokenize(text)
```

```
    postags = nltk.pos_tag(tokens)
```

```
    for j in range(len(postags)):
```

```
        if postags[j][1] == "VB" or postags[j][1] == "VBD" or postags[j][1] == "VBG" or postags[j][1] == "VBN" or  
postags[j][1] == "VBP" or postags[j][1] == "VBZ":
```

```
            cnt_vrb+=1
```

```
    return (cnt_vrb)
```

```
def adjectiveCount(text):
```

```
    cnt_adjctv=0
```

```
    tokens = nltk.word_tokenize(text)
```



```
postags = nltk.pos_tag(tokens)

for j in range(len(postags)):

    if postags[j][1] == "JJ" or postags[j][1] == "JJR" or postags[j][1] == "JJS":

        cnt_adjctv+=1

return (cnt_adjctv)

def adverbCount(text):

    cnt_adverb=0

    tokens = nltk.word_tokenize(text)

    postags = nltk.pos_tag(tokens)

    for j in range(len(postags)):

        if postags[j][1] == "RB" or postags[j][1] == "RBR" or postags[j][1] == "RBS":

            cnt_adverb+=1

    return (cnt_adverb)

def longWordCount(text):

    cnt_g5=0

    tokens = nltk.word_tokenize(text)

    for words in tokens:

        if len(words)>5:

            cnt_g5+=1

    return (cnt_g5)

def commaCount(text):

    cnt_comma=0

    tokens = nltk.word_tokenize(text)

    for comma in tokens:

        if comma==",":

            cnt_comma+=1

    return (cnt_comma)

def punctCount(text):

    cnt_punct=0
```

```
tokens = nltk.word_tokenize(text)

Punct = re.compile('[^A-Za-z0-9/-.]*')

filtered = [words for words in tokens if Punct.match(words)]

cnt_punct+=len(filtered)

return (cnt_punct)

def lexDivCount(text):

    tokens = nltk.word_tokenize(text)

    return (float(len(set(tokens))/len(tokens)))

def quoteCount(text):

    cnt_qt=0

    tokens = nltk.word_tokenize(text)

    for quote in tokens:

        if "" in quote:

            cnt_qt+=1

    return (cnt_qt)

def keyWordsCount(text):

    cnt_comma=0

    tokens = set(nltk.word_tokenize(text))

    for comma in tokens:

        if comma=="mri":

            cnt_comma+=1

    return (cnt_comma)

def keyWordsCount2(text):

    cnt_comma=0

    tokens = set(nltk.word_tokenize(text))

    for comma in tokens:

        if comma=="ct":

            cnt_comma+=1

    return (cnt_comma)
```

```
def keyWordsCount3(text):  
    cnt_comma=0  
    tokens = set(nltk.word_tokenize(text))  
    for comma in tokens:  
        if comma=="blood":  
            cnt_comma+=1  
    return (cnt_comma)
```

```
def keyWordsCount4(text):  
    cnt_comma=0  
    tokens = set(nltk.word_tokenize(text))  
    for comma in tokens:  
        if comma=="ecg":  
            cnt_comma+=1  
    return (cnt_comma)
```

```
for i in range(len(clean)):  
    wc.append(wordCount(clean[i]))  
    sc.append(sentCount(clean[i]))  
    nc.append(nounCount(clean[i]))  
    vc.append(verbCount(clean[i]))  
    adjc.append(adjectiveCount(clean[i]))  
    advc.append(adverbCount(clean[i]))  
    lwc.append(longWordCount(clean[i]))  
    cc.append(commaCount(clean[i]))  
    pc.append(punctCount(clean[i]))  
    kc1.append(keyWordsCount(clean[i]))  
    qc.append(quoteCount(clean[i]))  
    kc2.append(keyWordsCount(clean[i]))  
    kc3.append(keyWordsCount(clean[i]))
```

```
kc4.append(keyWordsCount(clean[i]))

dataset=[]

dataset.append(wc)

dataset.append(sc)

dataset.append(nc)

dataset.append(vc)

dataset.append(adjc)

dataset.append(advc)

dataset.append(lwc)

dataset.append(cc)

dataset.append(pc)

dataset.append(qc)

dataset.append(kc1)

dataset.append(kc2)

dataset.append(kc3)

dataset.append(kc4)

#%%% this is for svm/svr, you don't need to run this part

#this part also requires you to download the package called 'sk_learn'

dataset=np.transpose(np.array(dataset))

Xtrain, Xvalid, Ytrain, Yvalid = train_test_split(dataset, resolved_score, test_size=.1, random_state=66)

clf_svm = svm.SVC(C=10, kernel='linear', degree=2)

clf_svm.fit(Xtrain, Ytrain)

a = clf_svm.predict(Xvalid)

cohen_kappa_score(a, Yvalid, weights="quadratic")
```

Appendix H

Ethics Application and Approval for Amendment to Study

Amendment/Renewal to Study has been Approved

Amendment/Renewal ID: [Pro00071611_AME1](#)
Study ID: [MS2_Pro00071611](#)
Study Title: Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education
Study Investigator: [Tracey Stephen](#)

Description: The amendment/renewal to the above study has been approved.
Click on the link(s) above to navigate to the HERO workspace.
Please do not reply to this message. This is a system-generated email that cannot receive replies.

University of Alberta
Edmonton Alberta
Canada T6G 2E1

© 2008 University of Alberta
[Contact Us](#) | [Privacy Policy](#) | [City of Edmonton](#)

Appendix I

Permission Letter to Vice Dean, Faculty of Nursing to Apply AES Model

Tracey Stephen
tcs@ualberta.ca

June 29, 2018

Dr. Olive Yonge, Professor and Vice Dean
Faculty of Nursing, University of Alberta

Dear Dr. Olive Yonge

I am currently completing my doctoral studies in Educational Psychology with a specialization in Measurement, Cognition, and Evaluation in the Faculty of Education at the University of Alberta. My research area is focused on using technology to score responses to short answer exam questions. I am currently working on a research study titled, "Using Automated Essay Scoring to Assess Higher-Level Thinking Skills in Nursing Education".

The research project I am working on requires analysis of multiple responses to short answer questions. These responses are input to a computer software program that automatically scores the responses. It is based on the same notion of computers scoring multiple-choice questions (scantron) only this project focuses on computers scoring short answer type questions.

As you are aware, it is challenging to assess higher-level thinking skills in nursing students using *only* multiple-choice exams. In April 2016, three teaching teams at the Faculty of Nursing incorporated short essay questions into the exams that were administered to students. The students were given a patient scenario then asked 4 questions about the scenario. The incorporation of these questions was done in response to looking at ways to assess critical thinking in addition to multiple-choice exams. These exams were administered through *eclass* and all the responses were recorded and stored. I received ethics approval from the University of Alberta Research Ethics Board to run this study and also received permission from Dr. Joanne Profetto-McGrath to access this data in March 2017. This data has been used to build and score AES scoring models for the 4 part question and analyzed as outlined in my research study.

As a subsequent part for my study, I am requesting to access the 40 responses that FON students completed on June 14, 2018. All of these items were scored by a human and the students have received their final scores and course grades. I am wanting to do a follow up analysis of the AES model previously built from the initial data set and apply it to the new set of 40 responses. This study will have no impact on the students' scores or grades and all student identifying data will be removed from the responses so that only the actual responses are accessed. There will be no possible way to access any student identifying information. I have received ethics approval for the amendment to run the AES model on the additional 40 student responses and have attached this below.

I have attached a copy of my proposal for you to review. Thank you for your consideration of this important study affecting future nursing education and the use of technology for assessment of learning.


Regards,

Tracey

Tracey Stephen MN PhD (c) RN

(780) 492 3776

tcs@ualberta.ca

Appendix I cont'd**Approval Letter from the Vice Dean, Faculty of Nursing to Apply AES Model**Permission to access additional data  Inbox x**Olive Yonge**to me, Wendy 

Hello Tracey!

I am granting access to the 40 responses nursing students completed on June 14, 2018.

We have no secretarial support in the FoN office until July 9. Thus please use this email as your 'approval letter'.

Warmly,

Olive