HICKS, JUANITA C., Ph.D. The Use of Process Data to Examine Reading Strategies. (2019)
Directed by Dr. John T. Willse. 110 pp.

Researchers are increasingly interested in the cognitive behaviors students display during tests. This interest has led researchers to look for innovative ways to collect this type of data. Due to the proliferation of computer-based assessments, process data has become popular for its ability to help show what students know, what students don't know, and how students interact during assessments.

Aim: The aims of the current study are 1) to use process data to identify potential reading strategies and 2) to examine if reading strategy is associated with gender, race/ethnicity, and differences in performance.

Methods: Apply latent profile analysis (LPA) to extracted process data variables collected from US examinees who participated in the literacy section of the Program for the International Assessment of Adult Competencies (PIAAC). The variables are item response time and number of highlight events per item.

Results: A two-class solution provided the best fit for the data in each testlet of the literacy section of the PIAAC. Class one progressed through items in each testlet faster than class two. Class one most closely resembled a skimming strategy while class two most closely resembled a full-reading strategy. However, there was not conclusive evidence to suggest that the classes were reminiscent of skimming and full-reading. Class assignment had no significant relationship with gender nor race/ethnicity, and there was no significant difference in literacy performance between the two classes, except in one

case. Even then, both classes performed at a level two on the PIAAC literacy achievement scale.

Discussion: Response time was found to be the only discriminating variable in the identification of patterns related to reading strategies. While there was some separation between classes, it was minimal in some cases. Response time was found to be useful but not enough to identify conclusive reading strategies. Further research is needed to identify process data variables with explanatory power other than response time to aid in the identification of reading strategies.

THE USE OF PROCESS DATA TO EXAMINE READING STRATEGIES

by

Juanita C. Hicks

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2019

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation written by JUANITA C. HICKS has been approved by the following committee of the Faculty of The Graduate School at the University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____

Date of Acceptance by Committee

_____

Date of Final Oral Examination

ACKNOWLEDGMENTS

From the start to the end of this journey, my family has been supportive despite the geographical distance between us. So much of my motivation to finish this journey has come from my niece and two nephews, whom I wanted to show the possibilities of life, and remind them that it's not where you start but where you finish. I am also grateful to all my friends and fellow "GERMS" for helping me along the way. A special thank you to the "J-Squad": Julianne, Jeremy, and Justin for always trying to get me out of the house and have fun, and a special thank you to Cherie and Aileen for being my personal cheer team.

I cannot thank my committee members enough for their personal and professional guidance throughout my time in the department. Thank you, Dr. Luecht, for sharing with me your expertise in so many areas. Thank you, Dr. Henson, for being one of the best professors I have ever had. I truly admire your ability to reach students with different knowledge levels, at the same time. Thank you, Dr. Sunnassee, for your early guidance and continued support. Finally, thank you, Dr. Willse for allowing me to carve my own, slightly different path to this degree. I am thankful that I had the opportunity to learn from each of you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Process Data

The use of process data in educational assessment and measurement is growing at a substantial rate due to the proliferation of computer- and digitally-based assessments (Ercikan, 2018). Changes have been made in the educational landscape, and the latest technology has been employed to match these changes and deliver new assessments to students. Computer-based assessments log or track large amounts of data that could provide insight into the behavioral processes of students (Lee & Haberman, 2016). With this new data, researchers can now go beyond the information that test scores provide and begin to tap into other processes.

Process data can be collected from log files, key strokes, response times, eye-tracking, or the use of digital resources and other tools. The amount of knowledge that can be obtained from these different sources is substantial. Perhaps the most influential aspect of the use of process data has to do with the potential to help researchers understand what students know, what students don't know, and how students interact with test items (Ercikan, 2018). Performance profiles can be constructed for individual students or groups of students that provide information above and beyond test scores. Using process data to supplement test scores can help researchers determine how students are interacting with tests and if different student processes help or hurt those students'

overall performances. However, gathering this new information does pose challenges. New methodologies must be employed to accurately and appropriately collect, analyze, and interpret this complex data. Still, the unique possibilities and capabilities of the use of process data outweigh the challenges of this new data collection (Ercikan, 2018).

Reading Strategies

Process data gathered from computer- or digitally-based reading assessments can help aid in the observation and interpretation of behaviors used during tests. Specifically, these behaviors could be attributed to different strategies. Researchers are working to collect more data about skills and learning strategies to assess how students learn (Zhu, Shu, & von Davier, 2016) and to better provide instruction to students who aren't performing as well. Recently, research on reading strategies has come from eye-tracking studies (Hyönä, Lorch, & Kaakinen, 2002; van der Schoot, Vasbinder, Horsley, & van Lieshout, 2008; Duggan & Payne, 2011; Biedert, Hees, Dengel, & Buscher, 2012; Prichard & Atkins, 2016). Eye-tracking technology follows the direction of readers' eye movements, and logs the duration, location, and order of eye movements. Data from these eye-tracking studies suggest that there are certain behaviors or groups of behaviors that could represent features of reading strategies. Some of the most prominent reading strategies to be identified in the literacy literature are full reading (reading for comprehension or understanding) and skim-reading (Biedert et al., 2012; Duggan & Payne, 2011; Huddleston & Lowe, 2014), look-backs (Garner, Hare, Alexander, Haynes, & Winograd, 1984), distinguishing between important and non-important words (van der Schoot et al., 2008), and previewing (Prichard & Atkins, 2016).

Changes in the delivery of reading materials has changed the use of these strategies over time. The growing rate of digital media has had a significant effect on reading such that screen-based reading has emerged as a behavior (Liu, 2005). This type of reading behavior is characterized by spending more time on browsing and scanning, keyword spotting, non-linear reading, and reading more selectively, while less time is spent on in-depth and concentrated reading (Liu, 2005). Some of these behaviors have been observed in readers who interact with computerized or digital reading assessments.

Screen-based reading has influenced some readers to use shortcuts; only reading "important" sections of the text or test items and skim reading the rest. Often, readers feel that the amount of information presented in a text outweighs the time available to read it (Duggan & Payne, 2011). Readers may employ a selective strategy where they omit reading words, paragraphs or even pages. In this sense, the primary goal of the reader is not to read the whole text for comprehension but to selectively obtain key words from questions/items and then scan the text to match those key words to answers (Huddleston & Lowe, 2002). Process data, such as (RT) per item, RT per passage, and readers' chosen sequence of items can be used to identify potential behaviors of these strategies.

Another behavior associated with skim reading is the notion of "satisficing." Skim readers who use satisficing spend more time reading the first half of each paragraph, skip over the second half of each paragraph, and then proceed to the next paragraph (Duggan & Payne, 2011). This strategy is used because information gain tends to be lower in the latter half of paragraphs. Thus, readers don't feel like there is much new information presented beyond the beginning sections of paragraphs. Click-stream data could be used to identify readers using satisficing. For example, click-stream data would indicate the

sections that students highlight when reading and where these sections are located within a text. This information would indicate where students perceive information is gained or lost within a passage.

Another type of reading strategy is the use of "look-backs." This strategy refers to the frequency and duration of looking back to part of a text that has already been read (Hyönä et al., 2002; Garner et al., 1984). Using process data, this occurrence can be described by the frequency with which students refer to an initial passage after already reading it and the duration (using response time) of the lookback to that initial passage.

Distinguishing between important (goal-relevant) and unimportant (goal-irrelevant) words in text is another reading strategy. Words or phrases in text that address specific test items would be considered important; whereas, the topic of the paragraph might not be important (van der Schoot et al., 2008). In terms of process data, specifically response time, it might be possible to observe the amount of time students spend on items that ask for specific information located in the text and the amount of time students spend on items that do not ask for specific information located in the text.

Finally, previewing text is essential to any reading strategy. Previewing is one of the key strategies of top-down reading. Previewing the title, section headings, and even images to identify key information before text is read can help readers (Prichard & Atkins, 2016). For example, previewing the length of a long text may prompt the reader to read more quickly and previewing section headings may lead the reader to look at more relevant sections and skip over others (Prichard & Atkins, 2016). Process data also can be utilized to identify behaviors associated with this strategy as it can be used to identify other reading strategies.

There is little guidance in the literature about how well different strategies can be identified using process data. However, the use of process data in uncovering and identifying behaviors associated with reading strategies could help readers and researchers better recognize the processes that help or hurt readers in reading and literacy assessments. Research suggests that using any of these strategies individually or in combination, aids in comprehension and performance but only if they are used strategically and purposefully (Duggan & Payne, 2011; Huddleston & Lowe, 2014; Hyönä et al., 2002; van der Schoot et al., 2008). While the goal is to help readers read more carefully, rather than trying to prevent readers from using shortcuts, a better approach would be to help readers understand how to effectively use these strategies to meet their needs (Huddleston & Lowe, 2014).

Clustering Techniques

One of the most widely used educational data mining (EDM) techniques is clustering because of its simplicity in categorizing data into groups to look for patterns and provide structure to data (Dutt, Ismail, & Herawan, 2017). EDM techniques are useful for processing large amounts of data. Traditional clustering approaches such as k-means use a distance or similarity measure to group "like" data into the same group to minimize the within-cluster variance and maximize the between-cluster variance. A similar approach is latent class (LC) clustering. However, contrary to traditional clustering methods, the LC clustering approach is model-based. In LC clustering, a statistical model is hypothesized for the population from which the data are obtained (Magidson & Vermunt, 2002a; Magidson & Vermunt, 2002b). Simply, there are

underlying groups (latent variable for cluster membership) thought to cause a person's values on the observed variables (Pastor, Barron, Miller, & Davis, 2007).

Clustering methods like K-means or LC clustering are viable options in exploring patterns in process data to examine behaviors related to reading strategies. K-means would provide structure to reading process data by combining data into "like" groups. These groups could be based on different reading strategies, provided the interpretation of the data in the groups and the groups overall reflected reading strategies. LC clustering would use a similar process, but the data would be modeled to belong to an underlying class, in this case a latent variable reflecting reading strategy. The most notable LC clustering methods are latent class analysis (LCA) and latent profile analysis (LPA). The only difference between the LCA and LPA is the variables used in the model. In LCA, the variables are categorical while the variables in LPA are continuous. There is also the possibility of using both categorical and continuous variables in the same model, which is called "mixed-variable modeling" or "mixed-mode modeling" (Oberski, 2016). While the process and outcome of the models are similar, the LC clustering method does provide rigorous advantages over traditional clustering methods like K-means (Magidson & Vermunt, 2002a; Magidson & Vermunt, 2002b; Fraley & Raftery, 1998). I believe LC clustering would be the best method to identify potential reading strategies using process data from reading assessments.

The proposed research hopes to add to the literacy literature by proposing a novel way to identify potential reading strategies, using latent class (LC) clustering on process data that is available in most computer- or digitally-based assessment contexts. Eye-tracking technology is useful to examine reading strategies but also very expensive and

cumbersome to use with large samples. Think-aloud protocols, survey responses, and other self-report measures of the use of reading strategies may also lack reliability (Prichard & Atkins, 2016). Thus, using LC clustering with process data provides an economical and objective alternative to examining reading strategies.

Research Questions

To address and extend this work, I will be guided by the following research questions using exploratory analyses:

1) Can latent class (LC) analysis uncover latent classes that correspond to specific reading strategies?

1a). How many latent classes describe the data?

1b). Do patterns in these classes correspond to previously identified strategies?

2) Do uncovered latent classes have a relationship with any prominent observed variables?

2a). What is the prevalence of male and female examinees in these classes?

2b). What is the prevalence of different race/ethnicity groups in these classes?

2c). Is there a difference in overall achievement by latent class?

CHAPTER II

LITERATURE REVIEW

The following review of the literature contains seven major sections: process data, reading strategies, mixture models, latent class analysis (LCA), latent profile analysis (LPA), model fit and class selection, and missing data. Each section will outline research conducted to review the individual topic and together these sections will provide a foundation for the proposed research.

Process Data

Computer- and digitally-based assessments allow for the collection of detailed data that captures how examinees progress through assessments (Lee & Haberman, 2016). This data is commonly called process data. The interest in the use of process data is twofold; it supplements item responses and can provide additional useful information beyond test scores. Process data can be thought of as a reflection of examinees' problem-solving processes (Zhu et al., 2016), the expression of cognitive states and behavior that mediate the relationship between the construct and item score (Goldhammer & Zehner, 2017), or "the mechanisms that underlie what examinees do, think, or feel, when interacting with, and responding to, the item or task and are responsible for generating observed test score variation" (Hubley & Zumbo, 2017, p.2). Like item and test scores, process data can be used to draw inferences about examinees and what they know (Goldhammer & Zehner, 2017).

Process data can be gathered using a range of techniques such as think-aloud protocols, eye-tracking, video-ethnography, and log files (Maddox, 2017). While each technique has its own advantages and disadvantages, researchers mostly access process data using log files (Hobert, Sao Pedro, Raziuddin, & Baker, 2013; Lee & Haberman, 2016; Zhu et al., 2016). Although rich and authentic test environments exist, they aren't always easy to use to measure examinee knowledge and skills. There are a few key issues with using data from these environments. First, the log data from these environments is very complex, and second, there is a lack of theory for parsing and aggregating log data in a meaningful way (Hobert et al., 2013).

Research regarding process data is still new and researchers using process data are still contending with the best ways to parse and use process data meaningfully. Response time (RT) is one of the easier pieces of process data to extract. Response time refers to the amount of time examinees spend on items or other sections of an assessment. RT has become one of the main focuses of process data research (van der Linden, 2009; Lee & Haberman, 2016; Wise & Kong, 2005; Molenaar, 2015; Kong, Wise, & Bhola, 2007). Analysis of RT has become popular and has been found to offer information about examinees potentially speeding through tests (Guo et al., 2016; Kong et al., 2007), potential test-taking strategies (Lee & Haberman, 2016), and item position, item type, and item difficulty (Wise & Kong, 2005). The proposed research will utilize RT per item as one of two main process data variables to identify reading strategies.

The second process data variable used in the study will be the number of highlight events per item. Research has shown that highlighting relevant text plays an important role in encoding and organizing information for readers (Li, Tseng, & Chen, 2016; Silver

& Kreiner, 1997). Li et al. (2016) argued that text highlighting serves three main purposes: encoding information for reading comprehension, attention focusing on useful information, and acts as a visual cue for quick referencing. Highlighting and underlining relevant text has been shown to produce favorable benefits for students who engage in these strategies effectively (Mahdavi & Azimi, 2012; Li et al., 2016). For these reasons, assessing how examinees interact with the assessment using text highlighting may help in the identification of reading strategies.

Reading Strategies

Afflerbach, Pearson, and Paris (2008) provide a distinction between reading skills and reading strategies. In the literacy field, the terms are used interchangeably but are rarely distinguished from each other. The term "skills" has been used most in curricula for teachers and students, while the term "strategies" refers to the cognitive aspects of processing information (Afflerbach et al., 2008). Another reason to use the term "strategies" is that examinees can be taught to use them effectively and strategically. Moving forward, I will use the term "strategies" because I want to focus on the intentional, cognitive aspects examinees use during reading assessments.

The growing rate of digital documents has changed the way people read. Thus, research regarding the impact of digital media on reading has been a subject of exploration for many researchers (Liu, 2005). One of the main critiques of the growth of digital reading is that younger generations lack the ability to read and understand deeply (Bikerts, 1994). The natural assumption is that the same lack of ability to read and understand deeply applies to examinees using a digital medium for reading assessments. Liu (2005) surveyed 113 adults, ranging in age from 30-45 years old on their digital

reading habits. Over 80% of respondents reported a greater amount of time spent on browsing and scanning and 78% reported that they read more selectively, especially when presented with an overwhelming amount of information. Shockingly, about 45% of respondents reported a decrease in in-depth and concentrated reading. These types of behaviors were categorized as "screen-based" reading. The growth of digital reading environments can influence readers, especially young readers, to utilize more of the screen-based reading behaviors, where the goal is to not read to understand deeply but do just enough browsing and scanning to meet reading goals.

Huddleston and Lowe (2014) found that some of the screen-based reading behaviors can also be observed during print reading. In their study of 10 fifth-graders, they found that students used another strategy, like skim reading, called the "search-and-destroy method." This strategy is used by locating key words from items and then searching passages or text that match those key words to answers. Two studies (Greaney, 2004; Heafner & Spooner, 2008) suggest that the search-and-destroy method is commonly used by students but also that it is used unproductively by students. Students in the study by Huddleston and Lowe (2014) reported two reasons for using the search-and-destroy method: they thought reading was unnecessary and they thought reading was difficult. They also noted that stronger readers reported using this method infrequently. From the literature on the search-and-destroy method, researchers have noted two things: strategy instruction needs to be explicit and teaching students how to use strategies like the search-and-destroy method effectively would be more of a help rather than preventing students from using strategies at all (Huddleston & Lowe, 2014; Greaney, 2004; Heafner & Spooner, 2008).

Another strategy found in literacy research is the use of "look-backs." This strategy is characterized by reading the text or passage first, looking at an item, and then referring to the passage again to answer the item(s) (Garner et al., 1984). Using a sample of 24 upper elementary and middle school students, researchers explicitly taught 12 students how to use the look-back strategy while the remaining students served as the control group and received no instruction. Students were then asked to complete a short reading assessment consisting of passages and accompanying items. Seventy-two percent of students in the look-back instruction group correctly used the strategy when needed compared to only 30% in the control group. Nearly 70% of students in the instruction group answered items correctly using the look-back strategy compared to only 22% in the control group (Garner et al., 1984). Results from this study suggest that explicit instruction in the use of reading strategies can help with student achievement, as mentioned in Huddleston and Lowe (2014), Greaney (2004), and Heafner and Spooner (2008).

In a study by Biedert et al. (2012) researchers used eye-tracking technology to measure the eye fixations of readers reading different texts. They constructed a classifier that distinguished between different eye movements to create two classes: full reading and skimming. Full reading was characterized by longer (in time duration) and more eye movements over individual words. Skim reading was characterized by shorter (in time duration) and fewer eye movements over individual words, especially in a linear fashion. Using machine learning, feature detection, and linear classification, eye-movement data was classified by the above features to best model full reading and skimming. The classes, full reading and skimming were detected from the data with high precision,

recall, and accuracy (Biedert et al., 2012). The full reading class used more time to read passages, had longer gazes on words in the passages, and had eye movements that almost covered the entirety of the passages, while the skim reading class used less time to read passages, had shorter gazes on words in the passages, and had eye movements that didn't cover the entirety of the passages (Biedert et al., 2012). The difference in time used between the two classes was interesting to note; slower readers (more time) were classified as full readers while fast readers (less time) were classified as skim readers. Future research could examine the distinction of other classes or subclasses like fast full readers and slow skim readers.

A reading strategy related to skimming is "satisficing." In this strategy, readers engage in text until information gain begins to drop and then move on to the next section (Reader & Payne, 2007). Readers engage in this behavior because they often feel like there is not enough time to read through everything presented to them (Duggan & Payne, 2011). Normally, readers who use the satisficing strategy will start skipping words, paragraphs, or even full pages, resulting in less time on reading. Duggan and Payne (2011) sampled 28 students from the University of Manchester and asked them to read two articles of approximately 11 pages each. Eye-tracking technology was used to follow the readers' eye movements. Results indicated that readers spent more time reading the first half of each paragraph, skipped the second half of the paragraph, and then moved on to the next paragraph. These results would suggest that readers engaged in satisficing. Using heatmap analysis, Pernice (2017) noted that readers often engage in an "F-shape" pattern when satisficing. In this type of reading, readers read horizontally along the top edge of the text (the F's top bar), followed by a shorter horizontal direction further down

the text (the F's lower bar), and then finish by vertically scaling along the left side of the text (the F's stem). Satisficing may be a viable reading strategy when the goal is to allocate time and attention to only the most important parts of lengthy text.

Researchers state that successful readers know how to build an effective mental model of text using strategic reading methods (van der Schoot et al., 2008). One of those strategic reading methods is to incorporate goal-relevant (important) rather than goal-irrelevant (unimportant) information into the model. A sample of 36 students, in grades five and six, were asked to read two texts. Important and unimportant words, phrases, and sentences to understand the accompanying comprehension questions were noted by the researchers before the texts were given to the students. Using eye-tracking technology, students' eye movements were recorded as they read the two texts. Results showed that eye-fixation duration was significantly longer on important text elements compared to unimportant text elements (van der Schoot, 2008).

A final reading strategy potentially used by readers is previewing. This strategy is useful for examining information other than the main text to identify key information. Previewing the title, section headings, images, and the length of the text could help with detecting key information and preparation for the text to come. Most research on the use of reading strategies has been done with L1 (English speaking) samples. Very little research on reading strategies, especially using eye-tracking technology, has been done with L2 (English as a second language) samples. Previewing might be an important and useful strategy for L2 readers who may lack linguistic background knowledge to process and comprehend text in English (Prichard & Atkins, 2016).

Prichard and Atkins (2016) conducted a study with 38 Japanese students tasked with reading an expository text in English. Results of the eye-tracking study showed that the students used very little previewing of the text before starting to read linearly. It was not clear from the study if the lack of previewing is the norm for all L2 readers or just this group. It would also be important to examine if L2 readers use the previewing strategy when reading texts in their native language.

While the preceding literature focused on specific reading strategies, Mokhtari and Reichard (2002) examined reading strategies on a larger scale. In their study, they used a self-report inventory called the Metacognitive Awareness of Reading Strategies Inventory (MARSI) to assess middle and high school students' awareness and use of reading strategies when reading academic material. The MARSI consisted of 30 items and was administered to 443 students in grades 6-12. Factor analysis of the item responses resulted in three factors: global reading strategies, problem-solving strategies, and support reading strategies. Researchers defined global reading strategies as those that set the stage for reading, problem-solving strategies as those aimed at solving problems when text becomes difficult to read, and support reading strategies as those aimed at providing support for reading. The resulting three factors could be overarching strategies that the more specific reading strategies could fit under. For example, previewing may be a global strategy used to set the stage for reading, while look-backs or satisficing could be problem-solving strategies used when text becomes difficult to read. Other literacy strategies used by readers are re-reading, grouping words, and adjusting reading speed (Hare, 1981). Although these reading strategies are used by both proficient and less

proficient readers, proficient readers use the strategies strategically (Anderson, 1991; Hare, 1981).

Previous research has demonstrated the differential use of reading strategies for proficient readers and poor readers. Denton et al. (2005) showed that there is also a differential use of reading strategies for gender, such that females reported higher use of reading strategies compared to males. Wu (2014) also argued that female students had more knowledge of metacognitive reading strategies compared to male students. While there has not been much relevant research investigating the use of reading strategies between racial/ethnic groups, I believe it can be an important area of research, given that it has already been shown that there is an achievement gap related to ethnic/racial groups. Therefore, I will examine the use of reading strategies by gender and racial/ethnic groups to assess if there are differences in reading strategy use.

The preceding section introduced prominent research on the use and identification of reading and literacy strategies. Reading strategies from the literacy field have been identified using survey analysis, eye-tracking techniques, think-aloud protocols, observations and interviews. The current study proposes the use of process data and latent class (LC) clustering as a new method to identify potential reading strategies.

Mixture Models

Probability models, such as mixture models, have been used for a long time and are not new (Banfield & Raftery, 1993; McLachlan & Basford, 1988; Scott & Symons, 1971). However, in recent years, there has been an increased interest in mixture models due to its capability of clustering large amounts of data (Fraley & Raftery, 1998).

Clustering methods range from heuristic models (i.e., agglomerative clustering) to statistical models (i.e., LPA). The focus of this section will be on statistical models.

Mixture models are a type of latent variable model used to recover hidden, underlying groups from observed data (Oberski, 2016; Pastor, Barron, Miller, & Davis, 2007). Mixture models are based on a finite mixture of distributions, in which each mixture component corresponds to a different cluster or subpopulation. The "mixture" is referring to the notion that the data are not sampled from a population that can be described by a single probability distribution. Instead, the data is sampled from a population composed of a mix of distributions, one for each cluster, with each clusters' distribution characterized by its own set of parameters (Pastor et al., 2007; McLachlan & Chang, 2004). Clustering is then done by assigning each observation to the cluster to which it most likely belongs, based on a *posterior* probability that is conditional on the selected model and its estimated parameters (Pastor et al., 2007; McLachlan & Chang, 2004). The underlying groups are thought to cause a person's response on an observed variable.

The mixture modeling approaches I will consider are latent class analysis (LCA) and latent profile analysis (LPA). Throughout the literature on mixture modeling, LCA and LPA are considered tools that extend from the basic LC (also latent class) analysis model. The LC analysis model is the same basic model for LCA and LPA with slight changes to the parameters that distinguish between the two techniques.

Latent Class Analysis

Models based on latent variables play an important role in the behavioral sciences and educational research. Most of what interests researchers is unobserved; therefore,

researchers rely on empirical data that may be attributed to unobserved, underlying traits or behaviors (Dayton, 1991). Lazarsfeld (1950) extended the work of using mixture models by setting the stage for the use of LC analysis. LC analysis became more applicable in practice thanks to work by Lazarsfeld and Henry (1968) and Goodman (1974) who developed an algorithm to obtain the maximum likelihood estimates of the model parameters to describe the latent class distributions (Vermunt & Magidson, 2004; Muthén, 2001). Wolfe (1970) was one of the first to make the connection between LC analysis and cluster analysis.

Starting in the 90s, there was a renewed interest in the use of LC analysis as a clustering technique because the model was not as burdensome once computers and software became more computationally efficient (Vermunt and Magidson, 2002). In the early stages, LC cluster analysis was primarily modeled with categorical variables, thus, expanding LC cluster analysis to what is now called latent class analysis (LCA) (Vermunt & Magidson, 2002).

Latent class anaysis is a technique that aims to recover hidden groups based on the means of categorical (often binary) variables from observed data (Oberski, 2016). Like traditional clustering techniques, the size and number of classes is not known *a priori*. The goal of LCA is to divide the observations into mutually exclusive groups, such that observed variables are unrelated to each other within class (local independence). Any association between observed variables is accounted for only by the presence of the latent class (Templin, 2006). Traditional clustering methods put observations into groups because they are similar or related. Each latent class has a distribution that is characterized by its own set of parameters; therefore, the parameters of

an LCA model differ across the classes which form the categories of the latent variable (Vermunt & Magidson, 2004; Vermunt & Magidson, 2002; Muthén, 2001).

Consider an example with three categorical variables, A, B, and C, each with three response categories (1-3). The LCA model for this example is represented by the following equation:

$$\pi_{abc}^{ABC} = \sum_{x} \pi_{x}^{X} \pi_{a}^{A|X} \pi_{b}^{B|X} \pi_{c}^{C|X},$$

where *X* is the latent class variable, π_x^X is the size of latent class *x*, and $\pi_a^{A|X}, \pi_b^{B|X}, \pi_c^{C|X}$ are the probabilities that the variables (A, B and C) take on the response values (a, b and c) in the latent class *x*. The summation term states that the sum of the probabilities of any response value for any variable must sum to one across all latent classes (Oberski, 2016). In a more practical example, assume 100 students have taken a 25-item assessment. Responses to items are either correct (1) or incorrect (0). Using LCA, we might discover two classes; one class were the members have mastered the items and item means would be fairly high, and another class where the members have not mastered the items and item means are low (Dayton, 1991). The interpretability of the two classes suggest that the model may work to examine the achievement of the sample of students on the assessment.

Sainsbury and Benton (2011) conducted a study on the development of an e-assessment designed to give descriptive feedback on children's early reading skills. For the feedback to be effective it needed to be descriptive rather than numerical. Therefore, only using scores from the assessment wouldn't be enough. LCA was a viable option for

this study because it could generate information about patterns of performance, resulting in different subgroups, and the interpretability of those subgroups could aid in effective feedback.

Two parallel tests of 11 items each were created to address a range of early reading skills. The first five items assessed the identification of discrete consonant and vowel sounds. The next two items assessed recognition of rhymes. The following three items assessed word recognition, and the final items assessed sentence reading. The sample included 607 early elementary students, ranging in age from five to seven years old. LCA was used to explore the patterns of correct and incorrect responses to the 11 dichotomous items on both assessments. Models were run using two, three, four, and five classes. Researchers used the BIC and the Lo-Mendell-Rubin (LMR) likelihood ratio test of model fit (measures of model fit that will be explained in a subsequent section) to select the optimal number of classes. The four-class model was chosen as the best fitting model. The four-class model on both tests (same sample) resulted in similar results. The first class was labeled "sight" because students in this group performed poorly on most items but performed very well on the three word recognition items. The second class was labeled "sound" because students performed well on the five items related to consonant and vowel sounds. The third class was labeled "developing" because students performed fairly well on all items except the rhyming items. Finally, the fourth class labeled "balanced" because students performed well on all items (Sainsbury & Benton, 2011).

The reason LCA was used for this study was its capability of providing descriptive feedback about children's early reading skills. The descriptions of the classes suggest that children's early reading performance can be classified into meaningful

subgroups and this information can be used for descriptive feedback. Furthermore, the average scale scores for each class were similar at 23, 24, 24, and 30 for latent class one, two, three, and four, respectively. These findings show the value of LCA to capture fine-grained details of children's early reading skills above and beyond achievement scores alone. The findings from the above studies suggest the usefulness of LCA to produce meaningful, and interpretable subgroups from larger populations of interest.

LCA with Covariates

The following three sections will briefly describe models within the LCA framework. As the use of latent class models increases so does the intention of adding covariates to the model, so that the relationship between the covariates and latent class membership can be directly estimated. Some of the early researchers to propose the use of categorical covariates with latent class (LC) models were Clogg (1981), Goodman (1974), and Haberman (1979). This work then expanded to include continuous covariates (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Dayton and Macready, 1988; van der Heijden, Dessens, & Bockenholt 1996).

Muthén (2001) describes two cases where covariates are included in a latent class model. When variables serve as indicators of the latent class and covariates are included in the model to predict the latent class we can use the basic LC cluster model with added parameters for the covariates:

$$f(\boldsymbol{y_i}|\boldsymbol{z_i}, \theta) = \sum_{k=1}^{K} \pi_{k|\boldsymbol{z_i}} \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk}),$$

where K is the number of classes, f_k is the distribution of each class, $\boldsymbol{y_i}$ denotes scores on observed variables, θ_k describes the distribution parameters (μ_k, σ_k^2), J is the total number of indicators, j is a particular indicator, $\boldsymbol{z_i}$ denotes object i's covariate values, π_k is the probability of belonging to a class or the size of the class, and the appropriate univariate or multivariate (for sets of indicators) distribution for each element y_{ij} of $\boldsymbol{y_i}$ are be specified. When covariates have a direct effect on the indicator, the model becomes:

$$f(\boldsymbol{y_i}|\boldsymbol{z_i}, \theta) = \sum_{k=1}^{K} \pi_{k|\boldsymbol{z_i}} \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{z_i}, \theta_{jk}).$$

Latent class analysis (LCA) usually follows three steps: 1) the model is built for the data, 2) observed data are assigned to classes, and 3) the association between class membership and external variables like gender or race/ethnicity is examined using cross tabulations and chi-square analysis (Vermunt, 2010). However, this approach may underestimate the relationship between covariates and class membership. Currently, there are two common approaches to latent class modeling with covariates, the one- and three-step approach (Vermunt, 2010). The one-step method simultaneously estimates the LC model with the indicator variables and covariates. Most software packages can implement the one-step method. Although the three-step method may underestimate the relationship between the covariates and the latent classis it is usually the more practical approach, especially in an exploratory study.

Collins and Lanza (2010) propose using the step-wise approach to include covariates in latent class models. They argue that if the purpose of your research is to

identify characteristics that predict latent class membership then covariates should be included in the model but in the best way possible. First, baseline models should be run without covariates. The fit of the models and interpretability of the classes should then be evaluated, and if there are no major issues, then covariates may be included in the model. In the proposed research, described later, there are no strong *a priori* assumptions made about covariates. Therefore, I will be using the step-wise approach in my research.

There are two things to consider when running models that include covariates: missing data and covariate measurement scales. Most software does not allow missing data on covariates and will remove those with missing data. There are two things you can do with missing data on covariates: 1) remove cases with missing covariate data before running the baseline model so that the same data will be used in the baseline model and model with covariates, and 2) if removing cases causes severe data lost, multiple imputation may be used (Collins & Lanza, 2010). Regarding covariate measurement scales, covariates are treated as numerical; therefore, categorical covariates like race/ethnicity must be dummy coded before inclusion in the model. If interval or ratio scale covariates are used, these variables should be standardized before they are included in the model, especially if several covariates are to be included (Collins & Lanza, 2010). When the covariates are standardized, a one-unit change translates to a one-standard-deviation change for each variable, making it easier to compare effects across covariates (Collins & Lanza, 2010).

Confirmatory LCA

Until this point, the focus of LCA has been in an exploratory context where no specific assumptions about the number of classes are made *a priori*. Confirmatory latent

class analysis (CLCA) provides researchers with a tool for evaluating specific hypothesis about variables based on theory. CLCA is an approach where hypotheses about the number of classes as well as parameter constraints, according to theory, can be tested (Finch & Bronk, 2011). Goodman (1974) was one of the early researchers to introduce confirmatory latent class analysis.

Finch and Bronk (2011) state that three types of parameter constraints can be tested in CLCA. First, you can make equality restrictions, where item parameter values are constrained to be equal across latent classes. Next, you can set deterministic restrictions to test whether conditional response probabilities for a class equal 0 or 1. Finally, you can make inequality restrictions to test the likelihood of a latent class endorsing an item. Researchers can use one or all of the above strategies when testing CLCA models, but the selection of strategies should be supported by theory. Likewise, you can use CLCA to test a specific number of latent classes. For example, if theory from consumer research states that there are three types of shoppers, then you would test this result with a three-class confirmatory CLCA model. Model fit statistics such as AIC, BIC, entropy, and the Lo-Mendell-Rubin (LMR) likelihood ratio test would be used to evaluate the fit of your model and whether it follows the theory provided.

Schwartz and Zamboanga (2008) gave a practical example for the use of CLCA. In their study, they examined the extent to which Berry's (1997) acculturation categories emerged from a latent class analysis. Berry (1997) positioned acculturation as assimilation, separation, integration (biculturalism), and marginalization. Schwartz and Zamboanga (2008) had a sample of 436 Hispanic students enrolled in a large university in Miami. Thirty-four percent of the respondents were first-generation immigrants and

64% were second-generation immigrants. The American (15 items) and heritage (17 items) cultural orientation subscales of the Stephenson Multigroup Acculturation Scale were used as the primary indicators.

A combination of fit statistics and interpretability was used to settle on a six-class solution. Class one was labeled "undifferentiated," class two was labeled "assimilation," class three was labeled "partial biculturalism," class four was labeled "American-oriented biculturalism," class five was labeled "separation," and class six was labeled "full biculturalism" (Schwartz & Zamboanga, 2008). Results showed partial support for Berry's model. Six rather than four classes emerged from the LCA, and one class appeared to be a combination of Berry's original assimilation and integration clusters. However, three of Berry's acculturation clusters emerged from the LCA and multiple biculturalism classes were consistent with Berry and others' work (Schwartz & Zamboanga, 2008). One major difference is that Berry used K-means analysis to find his clusters while the Schwartz and Zamboanga (2008) study used LCA. K-means analysis does not have traditional fit statistics; therefore, it is not fully proven that Berry's acculturation clusters are correct. The proposed research will be conducted in an explanatory context so the use of CLCA will not be used. Future research on the identification of reading strategies using process data and LC clustering could lead to potential confirmatory latent class analysis models.

Latent Class Growth Analysis

Cross-sectional analysis focuses on a population or sample at one time point, while longitudinal analysis focuses on a population or sample over time. Researchers are increasingly interested in patterns of change across multiple time points (i.e., at least

three) (Andruff, Carraro, Thompson, & Gaudreau, 2009). Some of the reasons for the increase in the interest of latent class growth modeling (LCGM) and growth mixture modeling are the advances and availability of computer software to handle these analyses (Jung & Wickrama, 2008). The LCGM framework was extensively developed by Nagin (1999), who made substantial contributions to the theory and methodology of LGCM which was relatively new at the time.

A common approach to studying patterns over time is to use growth analysis, such as repeated measures analysis of variance (ANOVA) or latent growth modeling from the structural equation modeling (SEM) framework. Latent class growth modeling is a semi-parametric technique used to identify subgroups following a change in pattern over time (Andruff et al., 2009). Standard latent growth models estimate random coefficients for individual differences in the slope and intercept, while LCGM fixes the slope and intercept to be equal across individuals within a class (Andruff et al., 2009).

Extensions of LCGM would include estimating trajectories based on covariates and estimating a turning point or intervention that would change the developmental trajectory of the subgroups (Andruff et al., 2009). The strength of latent class growth models is that they are better estimated with multiple time points, so that a true trajectory is established. However, multiple time points can lead to greater attrition which can weaken the precision of the parameter estimates (Andruff et al., 2009). Therefore, researchers should make sure they have an adequate sample size and multiple time points to properly estimate parameters in a latent class growth model. The proposed research only focuses on a sample at one time point and will not be using LCGM. However, future

research could address the use of reading strategies over time using longitudinal data and latent class growth models.

Latent Profile Analysis

In recent years, there has been renewed interest in latent class (LC) cluster analysis using continuous variables. Latent profile analysis (LPA) is an extension of latent class analysis, where the former deals with continuous variables and the latter deals with categorical variables. Both are exploratory techniques. Lazarsfeld and Henry (1968) is mostly credited with the beginning stages of LC analysis which eventually set the stage for LCA, but Gibson (1959) conducted some early work on LC analysis with continuous variables which became known as latent profile analysis (LPA).

The basic LC cluster model for continuous variables has the following form:

$$f(\boldsymbol{y_i}|\theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y_i}|\theta_k),$$

where $\boldsymbol{y_i}$ denotes scores on observed variables, K is the number of classes, f_k is the density function for the distribution of each class, θ_k describes the distribution parameters (μ_k, σ_k^2) for each class, and π_k is the probability of belonging to a class or the size of the class (Vermunt & Magidson, 2002). To help with model convergence, the class variances are constrained to be equal, while the class means are allowed to vary and are freely estimated. Thus, data points are assigned to class based on their posterior probabilities (Muthén & Muthén, 2001). The two assumptions for a LPA model are local independence and normally or multivariate normally distributed variables (Templin,

2006). However, these are not strict assumptions. Input variables can be skewed or not normally distributed. In fact, skewness is sometimes expected in mixture models and can help determine classes (McLachlan & Peel, 2000; Muthén & Muthén, 2010).

Consider an example with a single continuous variable (y_i) that indicates cluster membership for person *i* and two clusters (*k*=2). A unique set of parameters for each cluster can be estimated. Parameters μ_1 and σ_1^2 can be estimated for cluster 1, and μ_2 and σ_2^2 can be estimated for cluster 2. These are called the distribution parameters (mean and variance). Mixing proportions or weights can also be estimated for each cluster, π_1 and π_2 for cluster 1 and cluster 2, respectively. These weights are constrained to be non-negative and must sum to one. The model for this example is then represented using the following equation:

$$f(y_i|\theta) = \pi_1 f_1(y_i|\mu_1, \sigma_1^2) + \pi_2 f_2(y_i|\mu_2, \sigma_2^2),$$

which shows that the distribution of the indicator variable (y_i), given the model parameters ($\theta = \pi_1, \mu_1, \sigma_1^2, \pi_2, \mu_2, \sigma_2^2$) is a weighted mixture of two separate distributions, each characterized by its own set of parameters (Pastor et al., 2007). This model can be easily expanded to fit more than two clusters by adding distribution parameters and a mixing proportion for each additional cluster.

In the case of multiple continuous variables, as in the multivariate case, the multivariate distribution of the *r* variables, contained in vector ($\boldsymbol{y}_i$), for person *i*, is considered to have a weighted mixture of *K* distributions. The multivariate representation of the above equation with *r* variables and *K* clusters is:

$$f(\boldsymbol{y}_i|\theta) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i|\boldsymbol{\mu}_k\boldsymbol{\Sigma}_k).$$

Again, the weights (π_k) are constrained to be non-negative and must sum to one. However, the distribution for each cluster *k* is now defined by a mean vector ($\boldsymbol{\mu}_k$), and a covariance matrix ($\boldsymbol{\Sigma}_k$), as opposed to a single mean and a single variance in the univariate case (Pastor et al., 2007).

The remainder of this section will describe practical research using latent profile analysis. Boscardin (2012) conducted a study using LPA to identify remedial students in medical schools. Currently, students are identified for remediation by using the Clinical Performance Examination (CPX). The CPX is a high stakes exam designed by clinicians and medical educators from all eight California medical schools. There is no universally accepted way to set the minimum competency, as the standards vary from school to school, and there is no consensus on the best methodological approach to identify students for remediation (Boscardin, 2012). The current method for identifying remedial students is problematic in two ways. The current cut-point of two standard deviations below the mean is not based on a theoretical framework and the cut-point is sample or cohort dependent. The performance distribution of the cohort can change the cut-point from year to year (Boulet, 2003).

Boscardin (2012) sampled 147 medical students from the University of California, San Francisco (UCSF) who completed the CPX exam in 2009. Latent class models with two, three, and four classes were fit to the data. BIC and entropy were used to assess model fit and choose the final model; the three-class model was selected as the

final model. The resulting three-class model represented three distinct groups. Class one and two were identified as students in need of remediation because their scores were lowest on two different sections of the CPX exam. Class three had the highest average score on all areas of the CPX exam (Boscardin, 2012). When comparing results of the LPA to the current policy, more remedial students were identified (targeting students at 1.5 standard deviations below the mean) increasing from 10% using the current method to 34% using LPA. In this study, the use of LPA as an alternative approach to identify remedial students in medical school resulted in a better identification of students in need compared to the current method.

Research on adolescent struggling readers (ASRs) has focused on reading proficiency rather than the challenges faced by ASRs (Brasseur-Hock, Hock, Kieffer, Biancarosa, & Deshler, 2011). Focusing on proficiency rather than challenges faced by ASRs puts a strain on those designing interventions for ASRs because there is no clear way to identify the areas that ASRs need the most help in. Brasseur-Hock et al. (2011) used LPA to identify and describe the subgroups of ASRs to better inform interventions and to provide targeted interventions to specific ASR subgroups. The goal of the study was to identify specific classes of ASRs and the component skill profiles they present using a two-step LPA. The sample included 319 students in the 9th grade from three urban high schools. Models were run with one, two, three, four, and five classes. Model fit was assessed using AIC, BIC, and distinct interpretability of the classes. The four-class model was chosen as the final model.

In order form lowest performing group to highest performing group on reading comprehension measures, the classes were labeled as “struggling comprehenders,” “low

average comprehenders," "average comprehenders," and "advanced comprehenders" (Brasseur-Hock et al., 2011). A LPA was fit to the combined sample of the lowest two comprehender groups, which was labeled "below-average comprehenders." A five-class model with this combined group was selected as the final model. The five subgroups within the "below-average comprehenders" group were, severe global weaknesses, moderate global weaknesses, dysfluent readers, weak language comprehenders, and weak reading comprehenders (Brasseur-Hock et al., 2011). The results of the LPA models identified subgroups of ASRs and specific weaknesses within the two lowest comprehender groups. Using this information, targeted interventions can be designed for adolescent struggling readers (ASRs).

Among students, the voluntary delay of tasks is not uncommon. According to Grunschel, Patrzek, and Fries (2013) there are usually two types of academic delay: 1) delay that is thoughtful and purposeful, and 2) delay that is irrational and potentially harmful. The second type is known as academic procrastination. As a whole, both types represent academic delay. However, researchers have focused on different types of academic procrastination to design interventions and help students who are most at risk (Grunschel et al., 2013). The current study used latent profile analysis (LPA) to explore different types of academic delayers to understand academic delay in general, and purposeful delay and academic procrastination specifically. The sample consisted of 554 college students enrolled in a range of subjects. Students' reasons for academic delay were assessed using a questionnaire of 14 items. However, not all 14 items were used as indicators in the LPA. Instead, a principal components analysis (PCA) was used to yield for main factors included in the LPA. The first factor was labeled "lack of study- and

self-management skills," the second factor was labeled "preference for pressure and past success," the third factor was labeled "worries and fears," and the last factor was labeled "discontent with studies."

Latent profile analysis resulted in a two-class model that provided clear classification of students. However, a four-class model offered more refinement and was selected as the final model using AIC, BIC, and the Lo-Mendell-Rubin (LMR) likelihood ratio test (Grunschel et al., 2011). About 94% of the students in the first class of the two-class model were split into the first and second class of the four-class solution. Eighty percent of the students in the second class of the two-class model were split into the third and fourth class of the four-class model. Class one had low expressions on all four factors and was labeled "inconspicuous," class two had the highest expression on pressure and past success and was labeled "successful pressure seeking," and class three and four had similarly high expressions on lack of study- and self-management skills. The "inconspicuous" group showed the lowest academic procrastination compared to the other three groups (Grunschel et al., 2011). On other measures such as the Big Five personality traits, the "inconspicuous" and "successful pressure seeking" groups had high conscientiousness scores. These results suggest that these two groups are purposeful delayers rather than academic delayers. The "worried/anxious" and "discontent with studies" groups endorsed high academic procrastination. Overall, these results show promise not only in academic procrastination and purposeful academic delay but also in the traits that make up these groups of students.

The preceding three studies using latent profile analysis (LPA) have been in an academic context. LPA is a viable option in almost all areas where the focus is to identify

patterns of behaviors in any context. The following research will focus on using LPA in the areas of health and criminology.

Adams et al. (2011) explored whether environment features were associated with adult physical activity and if this association produced distinct neighborhood profiles. Participants were recruited from Seattle, WA (n=1287) and Baltimore, MD (n=912). The Neighborhood Environment Walkability Scale (NEWS) was used to measure constructs such as transportation and urban planning related to physical activity. Variables from the NEWS were used as indicators in the LPA. Models with up to six classes were run for each sample. The Lo-Mendell-Rubin (LMR) likelihood ratio test and class interpretability were used to select the final model. For both samples, the four-class solution for neighborhood profiles was chosen. Class one was labeled "low walkable/low transit and recreationally sparse," class two was labeled "low walkable/recreationally sparse," class three was labeled "moderately walkable/recreationally dense," and class four was labeled "high walkable/ recreationally dense" (Adams et al., 2011). In both samples, more physical activity was documented in the "high walkable/recreationally dense" group. The main finding was that similar profiles were found in both regional samples which suggests generalizability of the profiles, although maybe only to major metropolitan areas within the U.S. Profiles were also associated with significant and meaningful differences in self-reported physical activity.

Vaughn, DeLisi, Beaver, and Howard (2008) state that typologies of burglars have been mostly based on qualitative data. Therefore, researchers wanted to construct a rigorous and methodological typology to identify burglars. The study sampled 456 adult, career burglars to identify burglar subtypes. Fifteen indicator variables that reflected a

range of offense characteristics were used. Five LPA models were examined, ranging from one to five classes. The final four-class solution was chosen based on BIC, entropy, and interpretability. Class one was the largest subgroup identified and was labeled "young versatile." This group did not have a criminal career span and had no unique characteristic offense pattern. Class two was labeled "vagrants" and had a high number of vagrancy offenses. Class three was labeled "drug-oriented burglars" and was characterized by a high number of drug possession offenses. Finally, class four was labeled "sexual predator burglars" and was characterized by a long criminal career span (i.e., more than 30 years) and numerous sexual offenses (Vaughn et al., 2008). The resulting burglar typologies were supported by literature; however, more works need to be done to explore other subtypes using larger data and other criminal offenses (Vaughn et al., 2008). The findings from the above studies support the use of LPA to identify patterns of data to produce significant, meaningful, and interpretable subgroups from larger populations of interest.

Mixed Mode Analysis

In the early stages of mixture modeling, the inclusion of both categorical and continuous variables was mathematically extensive (Lawrence & Krzanowski, 1996; Jorgensen & Hunt, 1996). Algorithms for the models were constructed to model categorical data or continuous data but not both simultaneously. Thus, latent class analysis (LCA) and latent profile analysis (LPA) had to be performed separately. However, in some cases, practical applications involve both categorical and continuous variables. Now, it is easier to include variables on different scales because software is more powerful (Vermunt & Magidson, 2002). Latent class (LC) cluster models can

handle nominal, ordinal, continuous, or count variables simultaneously. The general LC model becomes the following for mixed-mode data:

$$f(\boldsymbol{y_i}|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk}),$$

where $\boldsymbol{y_i}$ denotes scores on observed variables, K is the number of classes, f_k is the density function for each class, θ_{jk} describes the distribution parameters (e.g. μ_k, σ_k^2) for each class, J is the total number of indicators, j is a particular indicator, and π_k is the probability of belonging to a class or the size of the class. The appropriate univariate or multivariate (for sets of indicators) distribution for each element y_{ij} of $\boldsymbol{y_i}$ must be specified (Vermunt & Magidson, 2002). The proposed research could use a mixed mode model to assess the use of reading strategies if the indicator variables are on different scales.

Latent Class Clustering and Traditional Clustering Techniques

The premise of clustering is to partition a population of data into k sample sets. Traditional cluster analysis is often used to divide observed data into groups that share common characteristics (Boscardin, 2012). The most common traditional clustering technique is K-means clustering. MacQueen (1967) started early applications of the K-means clustering algorithm as similarity grouping for large data sets.

In this explanatory method, clusters are created such that the differences between clusters are maximized and the differences within clusters are minimized (Pastor et al., 2007). The K-means algorithm attempts to partition a given set of observations into

several clusters based on similarity, as to maximize the between-cluster variance and minimize the within-cluster variance. The process of maximizing between-cluster variance and minimizing within-cluster variance is performed iteratively using the following steps:

1. Randomly select "*c*" cluster centers, preferably as far away from each other as the data partition allows.
2. Calculate the distance between each data point and cluster centers; $|x_i - v_j|$, where x_i are the data points and v_j are the cluster centers.
3. Assign data points to the cluster center whose distance from the data point to the cluster center is smallest.
4. Recalculate new cluster centers using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{j=1}^{c_i} x_i$$

   where c_i represents the number of data points in the i^{th} cluster.
5. Recalculate the distance between each data point and the new obtained cluster centers.
6. Continue steps 3, 4, and 5 until cluster centers stop changing and data points are no longer reassigned.

The final objective function is to have the process iterate until the within cluster variance is at is minimum:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2,$$

where $\|x_i - v_j\|$ is the distance between the data points and cluster center, c_i is the number of data points in the i^{th} cluster, and c is the number of cluster centers (MacQueen, 1967). The K-means procedure can be sensitive to initial start values, so it may be best to try the method multiple times to see if different starting values produce widely different final clusters. If there is a lot of change in cluster generation based on different starting values, there may not be a "natural" clustering of the data (Rencher, 2012). K-means is easy to perform as well as computationally efficient, making it feasible to process very large data quickly into similar groups (Magidson & Vermung, 2002a; Magidson & Vermunt, 2002b).

Hierarchical methods of clustering are also considered traditional clustering techniques. These methods attempt to find "good" clusters in the data using computationally efficient techniques that involve sequential processes (Rencher, 2012). Established groups from these methods are also based on similarity. Hierarchical clustering has the distinct advantage that any valid measure of distance can be used. In fact, the observations themselves are not required; all that is needed is a matrix of distances (Rencher, 2012). The two main clustering approaches in hierarchical clustering are agglomerative and divisive clustering.

In each step of the agglomerative approach, an observation or cluster of observations is merged with another cluster, using a bottom-up approach. Each data point starts out as its own cluster, then the two nearest clusters (based on a distance measure

and linkage criteria) are merged into the same cluster. The process continues until there is only a single cluster, although it is possible to view cluster or observation merges at each step if visualized in a dendrogram (Rencher, 2012).

Some commonly used distance metrics for hierarchical clustering are Euclidean distance, Squared Euclidean distance, Manhattan distance, Maximum distance, and Mahalanobis distance. Each metric has its own formula for calculating the distance between two data points, but the simplest and most used metric is Euclidean distance (Rencher, 2012). The Euclidean distance between two points or two vectors of points (**a**, **b**) is calculated using the following formula:

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2},$$

which specifies a measure of dissimilarity between sets of observations. After the distance matrix has been calculated, linkage criteria determine the distance between sets of observations (similarity) as a function of the pairwise distances between observations (Rencher, 2012). There are several linkage criteria used with agglomerative clustering.

The single linkage method of clustering combines clusters by finding the "nearest neighbor" – the cluster closest to any given observation within the current cluster. The nearest neighbor method is defined by the minimum distance between a point in A and a point in B:

$$D(A, B) = \min\{d(y_i, y_j)\} \text{ for all } y_i \text{in A and } y_j \text{in B.}$$

The complete linkage method of clustering combines clusters by finding the "farthest neighbor" – the cluster farthest from any given observation within the current cluster. The farthest neighbor method is defined by the maximum distance between a point in A and a point in B:

$$D(A,B) = \max\{d(y_i, y_j)\} \text{ for all } y_i \text{in A and } y_j \text{in B.}$$

The average linkage method is similar to the single and complete linkage methods with the exception that the distance between clusters is now represented by the average distance of all objects within each cluster:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j),$$

where the sum is over all $y_i$in A and $y_j$in B.

In the centroid method, the distance between two clusters *A* and *B*, is defined as the Euclidean distance between the mean vectors (centroids) of the two clusters:

$$D(A,B) = d(\bar{\boldsymbol{y}}_{\boldsymbol{A}}, \bar{\boldsymbol{y}}_{\boldsymbol{B}}),$$

where $\bar{\boldsymbol{y}}_{\boldsymbol{A}}$and $\bar{\boldsymbol{y}}_{\boldsymbol{B}}$ are the centroids of cluster *A* and cluster *B*. After two clusters are joined, the centroid of the new cluster *AB* is defined by the weighted average:

$$\bar{\boldsymbol{y}}_{AB} = \frac{n_A \bar{\boldsymbol{y}}_{\boldsymbol{A}} + n_B \bar{\boldsymbol{y}}_{\boldsymbol{B}}}{n_A n_B}.$$

One issue with the centroid method is that the centroid is closer to the cluster with the most observations. To avoid the effect of sample size, the median method can be used. The median method uses the average of the group means (centroids) without considering sample size:

$$m_{AB} = \frac{1}{2}(\bar{\boldsymbol{y}}_{\boldsymbol{A}} + \bar{\boldsymbol{y}}_{\boldsymbol{B}}).$$

Finally, Ward's method attempts to minimize the increase in SSE (sum of within-cluster distance) when joining two clusters, *A* and *B*:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B).$$

The choice of which distance metric and linkage method to use is up to the researcher. However, different combinations of the two may have a big impact on the results so reasoning for which metric and linkage method used should be provided (Rencher, 2012).

Although divisive clustering is a hierarchical clustering approach, it works in the opposite direction of agglomerative clustering. In divisive clustering, data starts in a large cluster and is then divided into subgroups until there are as many clusters as individual observations. I will not go in to further detail regarding divisive clustering as its goal is not like latent class analysis (LCA), latent profile analysis (LPA), agglomerative clustering, or K-means clustering.

While latent class analysis (LCA), latent profile analysis (LPA), K-means clustering and agglomerative attempt to group data into meaningful clusters, there are some advantages to using LCA and LPA over K-means and agglomerative clustering.

LCA and LPA are model-based clustering techniques which means they use a statistical model to try to optimize the fit of the model to the data (Fraley & Raftery, 1998). An advantage of using a statistical model is that the number of clusters chosen is not arbitrary and is based on statistical tests and fit criteria, unlike tradition clustering methods. Also, decisions regarding scaling of the observed variables do not have to be made for the model-based clustering approaches. However, variables must be standardized to have equal variance prior to running methods like K-means to avoid clusters that are dominated by variables with the largest variance Magidson & Vermunt, 2002a; Magidson & Vermunt, 2002b). Model-based clustering can also deal with variables of mixed measurement scales and include covariates unlike traditional clustering methods. A model-based clustering method will be used in the proposed research because it provides distinct advantages over traditional clustering methods.

Model Fit and Class Selection

Choosing the "correct" number of clusters in a cluster analysis is difficult. Often, the number of clusters chosen is not necessarily correct, but does adequately fit the model (Celeux & Soromenho, 1996). Therefore, measures to help assess the fit of the model are needed to adequately interpret the model. Researchers usually use a combination of fit criteria to evaluate the number of classes/clusters in a model. The criteria fall into three categories: 1) information-theoretic models, 2) likelihood-ratio tests, and 3) entropy-based criterion (Tein, Coxe, & Cham, 2013).

The information-theoretic fit criteria are the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC has had a significant impact on the evaluation of statistical models. AIC is one of the most commonly cited and widely used

model fit statistics (Bozdogan, 1987). It was first introduced by Akaike (1973) and Akaike (1974). The formula to calculate the AIC is:

$$AIC = 2k - 2\ln(\hat{L}),$$

where k is the number of components, and $\hat{L}$ is the maximum value of the likelihood function for the model with k components (Celeux and Soromenho, 1996). The AIC is a measure of goodness of fit that only considers the number of model parameters, in this case k (McLachlan & Chang, 2004). The BIC is a goodness of fit measure that considers the number of observations in addition to the number of model parameters (McLachlan & Chang, 2004). It was first introduced by Schwarz (1978). The formula for the BIC is:

$$BIC = \ln(n)\, k - 2\ln(\hat{L}),$$

where k is the number of components, $\hat{L}$ is the maximum value of the likelihood function for the model with k components, and n is the sample size (Celeux and Soromenho, 1996). AIC and BIC have been the most commonly used criteria for assessing model fit and the number of components in mixture models (McLachlan & Peel, 2000). When evaluating models using AIC and BIC lower values of these criteria suggest better fit.

One of newer approaches to model fit is the likelihood ratio test. The bootstrapped likelihood ratio test (BLRT) (McLachlan, 1987) and the Lo-Mendell-Rubin (LMR) likelihood ratio test (Lo, Mendell, & Rubin, 2001) are formal significant tests of the null hypothesis that a k_0 component mixture fits better than an alternative hypothesis of k_1, where k_1 is greater than k_0. In the case of a mixture model or latent class clustering

model, the BLRT and LMR likelihood ratio tests are used to compare models with increasing numbers of classes. Smaller *p*-values on these tests provide support for the more complex model with *k* clusters compared to *k*-1 clusters (Pastor et al., 2007).

The last approach of model fit and class selection is entropy. Entropy is a measure of classification uncertainty and is bound between zero and one using the following equation (Muthén, 2000):

$$E_K = 1 - \{\Sigma_i \Sigma_k [-p_{ik} \log(p_{ik})] / n \log(K)\}.$$

There is no agreed upon value of entropy as "good enough"; however, entropy values closer to one suggest good classification and highly discriminating latent classes (Celeux & Soromenho, 1996). Although the use of these fit criteria are helpful in determining the number of classes in LC clustering, when the degree of separation between classes is small the fit criteria poorly select the correct number of classes (Tein et al., 2013). The use of these fit statistics along with interpretability of classes/clusters should help with choosing the best model for your data.

Missing Data

The case of missing data can be challenging for some. Often, there will be cases where data is missing on one or multiple variables. Most software can handle missing data with no issue under the missing at random (MAR) assumption (Muthén & Muthén, 2017; Collins & Lanza, 2010). The software Mplus (Muthén & Muthén, 1998-2017), uses all available data under full information maximum likelihood (FIML) estimation to handle missing data in latent class analysis (LCA) and latent profile analysis (LPA)

(Muthén & Muthén, 1999; Muthén & Muthén, 2017). In the proposed research, missing data is not an issue, even though the proposed model is robust enough to account for missing data.

Summary of Literature

The preceding sections of this literature review have presented research on process data, literacy, and methods on clustering to inform the proposed study. Reading strategies from the literacy field have been identified using survey analysis, eye-tracking techniques, think-aloud protocols, observations and interviews. Research regarding process data has shown a unique opportunity to provide information about students behavioral and cognitive processes during assessments that could be potentially related to reading strategies. Latent class (LC) cluster analysis techniques such as latent class analysis (LCA) and latent profile analysis (LPA) have been shown to be useful options for finding subgroups in data. The proposed research will use LC clustering to attempt to uncover behaviors related to reading strategies using process data.

CHAPTER III

METHODOLOGY

The following section outlines the methodology to be used for the proposed research. This study proposes a novel way to examine reading strategies by utilizing latent class (LC) analysis to identify subgroups (patterns) using process data from a computer-based assessment. The review of literature of studies using LC analysis has shown that it is a viable option for exploring patterns of data that reflect underlying subgroups, in the context of this study, potential reading strategies. The use of process data is proposed as the data of interest because of its potential to capture behavioral and cognitive processes of examinees. The captured processes could be related to the use of reading strategies during an assessment.

Two main research questions guide this study:

1) Can latent class (LC) analysis uncover latent classes that correspond to specific reading strategies?

1a). How many latent classes describe the data?

1b). Do patterns in these classes correspond to previously identified strategies?

2) Do uncovered latent classes have a relationship with any prominent observed variables?

2a). What is the prevalence of male and female examinees in these classes?

2b). What is the prevalence of different race/ethnicity groups in these classes?

2c). Is there a difference in overall achievement by latent class?

Data Source

The Program for the International Assessment of Adult Competencies (PIAAC) is a large-scale assessment developed by the Organization for Economic Cooperation and Development (OECD) (National Center for Education Statistics [NCES], 2018). PIAAC's goal is to assess and compare the basic skills and competencies of adults around the world. PIAAC is unique compared to other international assessments in that is the first large-scale, adaptive assessment administered by laptop to respondents in their homes.

The adaptive process means that respondents will be directed to a set of easier or more difficult items based on their answers to previous sections of the assessment. On the PIAAC, participants first answer the information communication technology (ICT) core section, which measures basic computer skills. If participants perform well on the ICT core they are directed to the literacy/numeracy core, which measures basic skills in these domains. If participants perform well on both core components of the PIAAC they are randomly routed to the computer-based literacy, computer-based numeracy, or problem-solving domains. For respondents receiving the literacy domain, they are routed to one of three panels of items (e.g. testlets) on stage one. Based on performance on stage one, respondents are then routed to one of four testlets, ranging from easy to more difficult on stage two (NCES, 2018).

The Program for the International Assessment of Adult Competencies (PIAAC) is designed to assess a range of abilities from simple reading to complex problem-solving in four domains: literacy, numeracy, problem-solving in technology rich environments (PS-

TRE), and reading (NCES, 2018). All participating countries are required to assess literacy and numeracy, while PS-TRE and reading are optional. I will be focusing on process data from the literacy domain. One of the goals of international assessment programs such as PIAAC is to assess reading literacy skills (Gil, Martinez, & Vidal-Abarca, 2015). The literacy domain of the PIAAC assessment was chosen, in the absence of reading data, because it serves as an adequate data source for the investigation of reading strategies. The literacy domain includes items containing continuous text (e.g. sentences and paragraphs), non-continuous texts (e.g. graphs and maps) and electronic texts (e.g. text in an interactive environment). Items in this domain are meant to assess three cognitive processes: access and identify, interpret and integrate, and evaluate and reflect (NCES, 2018). PIAAC's literacy domain assesses adequate processes that are related to the potential use of reading strategies.

There are six achievement levels for the PIAAC assessment: below level one, level one, level two, level three, level four, and level five. Each level has different task descriptions provided for what examinees should be able to do at each level (NCES, 2018). Scale score plausible values (PVs) are associated with each achievement level. Plausible values represent what the performance of an individual might have been had they taken the entire assessment. Each examinee has 10 PVs that are randomly drawn from an empirically derived distribution of score values based on students' item responses and background variables (NCES, 2018). To get a final scale score for each examinee, all 10 PVs have to be used together. In this study, the average of all 10 PVs for each examinee is used as the final scale score.

Log files for the Program for the International Assessment of Adult Competencies (PIAAC) can be collected because the assessment is administered via computer. During the assessment, user interactions are logged automatically. Most of the users' actions within the assessment tool are recorded and stored with time stamps in separate files (OECD, 2017). The log files contain data for each participant on all domains assessed. Data collection for the US was conducted between August 25, 2011 and April 3, 2012. All adults in private households between the ages of 16 and 65 that lived in the country during the time of data collection, regardless of citizenship, nationality or language were eligible for the assessment (OECD, 2017). This sampling strategy resulted in a total sample of 4,094 respondents from the US 2012 round of PIAAC.

Population and Sample

There were 4,094 total respondents in the U.S. in the first round of the Program for the International Assessment of Adult Competencies (PIAAC) in 2012. Log file data was collected for each respondent. There was a different sample size for each stage and testlet of the 2012 PIAAC because the assessment was adaptive and not all respondents took all items. Due to missing data, four cases were dropped from both stage one testlet one and testlet two, respectively, while one case was dropped from stage one testlet three. Two cases were dropped from stage two testlet one, one case dropped from stage two testlet two, two cases dropped from stage two testlet three, and no cases dropped from stage two testlet four. Therefore, literacy stage one testlet one, two, and three had sample sizes of 824, 871, and 945, respectively. Literacy stage two testlet one, two, three, and four had sample sizes of 501, 694, 689, and 758, respectively. These samples were obtained by only keeping cases with complete data. Tein et al. (2013) state that on the

low end a sample size of 250 is adequate to conduct LPA, but sample sizes closer to 1000 are better, especially if many variables are included in the model.

Research Design

The Program for the International Assessment of Adult Competencies (PIAAC) is an adaptive test. The required domains, literacy and numeracy, are given in stages and testlets made up of individual items. The literacy and numeracy domains are tested in two stages. Stage one has three testlets made up of nine items each while stage two has four testlets made up of 11 items each.

Stage one testlet one, two and three, have a total of 13, 14, and 16 process data variables, respectively. There are four highlight event variables on testlet one, five highlight event variables on testlet two, and seven highlight event variables on testlet three. Each testlet has nine RT variables (one for each item). Stage two testlet one, two, three, and four, have a total of 19, 18, 20, and 18 process data variables, respectively. There are eight highlight events on testlet one, seven highlight events on testlet two, nine highlight events on testlet three, and seven highlight events on testlet four. Each testlet in stage two has 11 RT variables, again, one for each item.

To address research question one, I will analyze variables extracted from the process data collected on items in each literacy testlet from stage one and two. The following variables will be extracted from the process data using the PIAAC LogDataAnalyzer (OECD, 2017): response time (RT) per item, and the number of highlight events per item. RT is used as a variable of interest based on its wide use in process data research and highlight events are included as a variable of interest because items in each testlet allow for users to highlight portions of text to help answer items. I

examined the distribution of the highlight events variable, based on the variances of responses for each item, to see if it was better used as a continuous or categorical variable. The highlight events variable was sufficient as a continuous variable, therefore a LPA (both variables are continuous) was used to analyze the patterns of the process data variables. Biedert et al. (2012) showed that full readers used more time when reading text while skim readers used less time when reading text. The use of RT as a variable in the proposed research will examine how examinees use their time (e.g. skimming or full reading). The highlight events variable will assess how examinees process supporting information when answering items. It is common practice to use the logarithmic transformation of response time to create a symmetric and normal distribution because of the tendency of RT to be positively skewed (van der Linden, 2009). However, no transformation of RT will be done because LPA can handle non-normal data, as the requirement for normally distributed data is not a strict assumption of LPA.

To address research question one, part a and b, I will evaluate model fit and select the number of latent classes using information criteria AIC and BIC, the Lo-Mendell-Rubin (LMR) and bootstrapped likelihood ratio tests, entropy, and interpretability of latent classes as described in the model fit and class selection section of the literature review. To evaluate fit I will focus on models with lower values of AIC and BIC and entropy values closer to one. The likelihood ratio tests will suggest models of increasing complexity based on significance tests; when significance is not met at an alpha level of .05 I will stop running models with increasing complexity. Overall, I will assess models using the best combination of these criteria. In some cases, the fit criteria will not agree and will suggest different models as the best fitting model. In this case, I will make a

judgment call based on practicality of model fit and interpretability of results that correspond to previously identified strategies.

To answer research question two, part a and b, I will use chi-square tests for associations analyses to examine the prevalence of males and females, and different race/ethnicity groups in the uncovered latent classes. For research question two, part c, I will perform significance tests to examine if there are any significant differences in the average achievement score between the uncovered latent classes.

Chapter Summary

Different methods have been used to assess and uncover the use of reading strategies on computer-based assessments; yet, none of have utilized latent class (LC) analysis using process data. This chapter proposed a new design that would use available data collected from computer-based assessments and a viable clustering technique to evaluate the use of reading strategies. This study adds to the body of literature by utilizing a new data source that has the potential to provide behavioral and cognitive information about examinees. This data can supplement score information to provide a well-rounded description (i.e., performance profile) for examinees or groups of examinees.

CHAPTER IV

RESULTS

Latent profile analyses were completed using the Mplus software, version 8 (Muthén & Muthén, 1998-2017) while chi-square tests, significance tests, and descriptive statistics were completed in SPSS, version 25 (IBM Corp, 2017). Results of the study are presented by stage and testlet. Due to variable name length restrictions, items were renamed using the RT or HI prefix to denote item level response time or highlight events, followed by a number indicating item sequence. Upon initial inspection of data for each testlet, there were significant outliers on RT variables. In some instances, respondents spent more than 25 minutes on a single item while spending less than two minutes each on the rest of the items. To prevent egregious outliers from skewing the results, the top five response times per item, in each testlet were examined. If there was a significant drop off in the response time between the top five entries, then respondent(s) with the highest response times would be dropped from the data. This examination resulted in dropping four respondents from stage one testlet one, three respondents from stage one testlet two, two respondents from stage one testlet three, one respondent from stage two testlet one, three respondents from stage two testlet three, and two respondents from stage two testlet four. No respondents were dropped from stage two testlet two.

Stage One Testlet One

Descriptive Statistics

Table 1 summarizes the descriptive statistics of select demographic characteristics chosen from the PIAAC background questionnaire. Stage one testlet one contained 820 respondents. There were 362 (44%) males and 458 (56%) females in the testlet sample. By race/ethnicity, there were 524 (64%) white, 91 (11%) Hispanic, 147 (18%) black, and 58 (7%) other respondents included in the sample. The majority of respondents also came from city 312 (38%) and suburban 220 (27%) locations. Nearly half of the sample was between the age of 16 and 34, while the other half of the sample ranged in age from 44 to 71 and above.

Table 1. Stage One Testlet One Demographic Variables

Gender	n	%	Age	n	%
M	362	44%	16-19	89	11%
F	458	56%	20-24	123	15%
Location	n	%	25-29	103	13%
City	312	38%	30-34	83	10%
Suburb	220	27%	35-39	62	8%
Town	88	11%	40-44	59	7%
Rural	200	24%	45-49	47	6%
Race	n	%	50-54	57	7%
Hispanic	91	11%	55-59	62	8%
White	524	64%	60-65	67	8%
Black	147	18%	65-70	45	5%
Other	58	7%	71+	23	3%

Table 2 presents the descriptive statistics for all quantitative variables: item response time (RT), item highlight (HI) events, and score. For response time variables, all statistics are in seconds while statistics for highlight events variables are in terms in

individual highlights. On average, respondents spent less than a minute on the items and had less than four highlight events. Respondents also had an average score of 269 on the literacy section, which is a level two on the achievement level scale.

Table 2. Stage One Testlet One Quantitative Variables

	RT1	RT2	HI2	RT3	RT4	RT5	HI5
Mean	39.28	44.95	1.57	36.90	29.07	71.63	2.58
SD	32.18	35.38	1.24	24.07	21.92	51.79	2.97
Min	2.69	1.14	0.00	1.22	2.38	2.17	0.00
Max	446.01	352.23	12.00	356.98	201.36	538.74	20.00
Skewness	4.61	3.25	2.53	6.31	3.23	2.49	2.67
Kurtosis	39.99	17.77	11.56	72.87	15.71	12.73	9.62

	RT6	RT7	RT8	HI8	RT9	HI9	AvgPV
Mean	31.57	59.31	60.99	3.53	41.99	3.55	269.39
SD	23.46	39.39	46.71	3.89	31.26	4.12	44.84
Min	1.87	1.99	3.61	0.00	2.00	0.00	135.21
Max	342.51	366.18	519.36	42.00	260.05	43.00	395.99
Skewness	4.85	2.05	3.02	3.75	2.46	4.43	-0.33
Kurtosis	45.30	7.76	16.44	22.75	9.42	29.33	-0.04

Final Model

Latent profile analysis for the data in stage one testlet one was conducted using a two-, three-, and four-class model. Information criteria and entropy values, along with the significance values for the Lo-Mendell-Rubin (LMR) likelihood ratio test (LRT) and bootstrapped likelihood ratio test (BLRT) for each class are in Table 3. The BLRT always favored the k-class solution over the k-1 class solution, while the LMR-LRT did not favor the k-class solution over the k-1 class solution, using an alpha level of 0.05. Likelihood ratio tests were inconclusive and did not show clear support for one model over another. Although AIC and BIC values decreased, while entropy values increased

with more classes, there was not enough information to distinguish classes from each other in the three- and four-class model (Figure 1 and Figure 2). Estimated means and variances for the three- and four- class model are provided in Table 4 and Table 5, respectively. The presence of additional classes resulted in sample sizes of less than 10. Therefore, based on interpretability, limited fit criteria support, and small sample sizes, in the additional classes, the two-class model was chosen as the final model (Figure 3). In all testlets, across both stages, there was not enough information to suggest a three- or four-class model as the best fitting model over the two-class model. Therefore, the visual class inspections and estimated means and variances are not shown beyond the two-class model for the following results.

Table 3. Stage One Testlet One Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	86866.36	87054.73	0.92	0.21	< 0.01
3-Class Solution	86328.90	86583.20	0.95	0.33	< 0.01
4-Class Solution	85828.41	86148.64	0.96	0.32	< 0.01

Figure 1. Stage One Testlet One Estimated Means of Process Data Variables (C3)

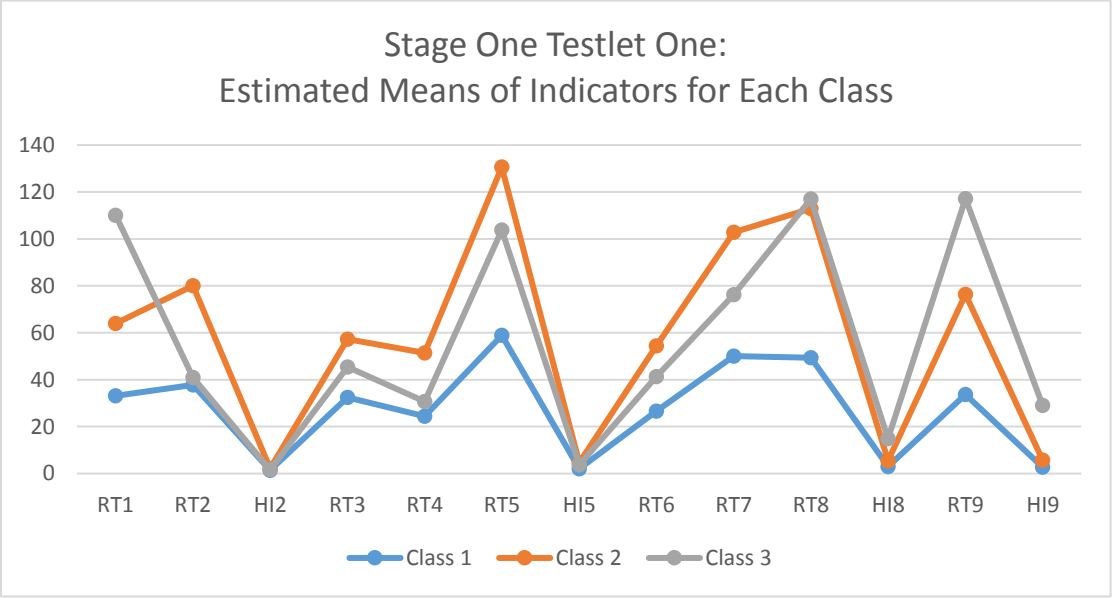

Table 4. Stage One Testlet One Estimated Means and Variances (C3)

	Class 1 (n = 674)		Class 2 (n = 136)		Class 3 (n = 10)	
Variable	Mean	Variance	Mean	Variance	Mean	Variance
RT1	33.09	838.03	64.02	838.03	110.12	838.03
RT2	37.73	998.20	80.06	998.20	40.87	998.20
HI2	1.44	1.45	2.19	1.45	1.60	1.45
RT3	32.53	491.69	57.30	491.69	45.49	491.69
RT4	24.40	377.13	51.45	377.13	30.65	377.13
RT5	58.92	1944.16	130.60	1944.16	103.86	1944.16
HI5	2.11	7.83	4.70	7.83	3.68	7.83
RT6	26.67	439.78	54.48	439.78	41.34	439.78
RT7	50.03	1154.51	102.81	1154.51	76.29	1154.51
RT8	49.38	1573.58	112.89	1573.58	117.01	1573.58
HI8	2.95	12.65	5.49	12.65	14.92	12.65
RT9	33.73	650.32	76.42	650.32	117.20	650.32
HI9	2.72	7.66	5.69	7.66	29.12	7.66

Figure 2. Stage One Testlet One Estimated Means of Process Data Variables (C4)

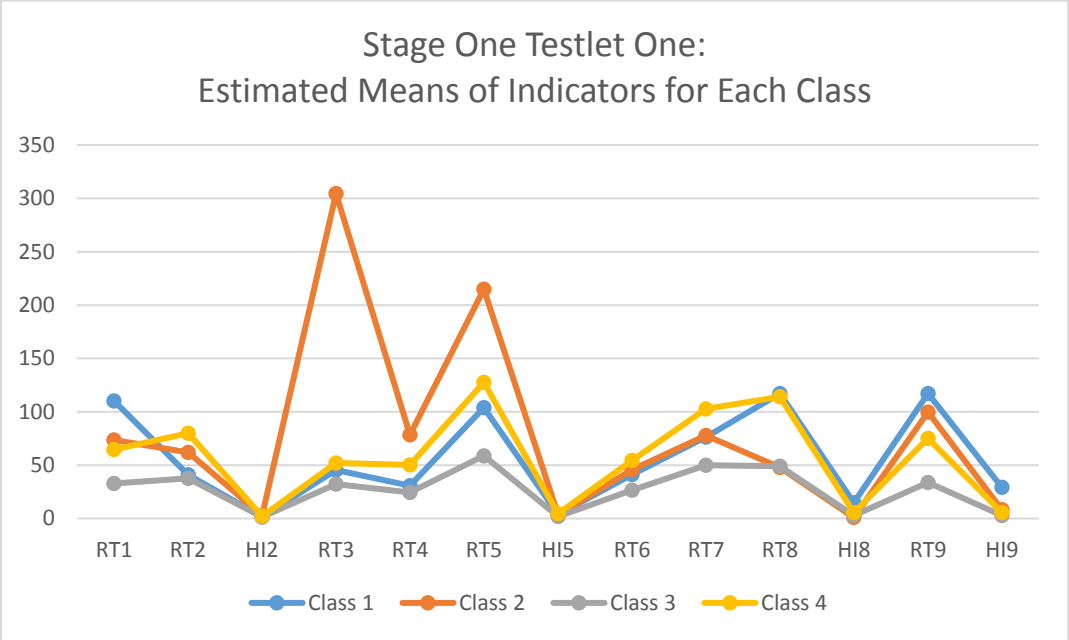

Table 5. Stage One Testlet One Estimated Means and Variances (C4)

	Class 1 (n = 10)		Class 2 (n = 3)		Class 3 (n = 673)		Class 4 (n = 134)	
Variable	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
RT1	110.12	831.62	73.78	831.62	32.91	831.62	64.52	831.62
RT2	40.87	998.24	62.01	998.24	37.63	998.24	79.82	998.24
HI2	1.60	1.44	3.00	1.44	1.43	1.44	2.17	1.44
RT3	45.49	259.51	304.51	259.51	32.36	259.51	52.15	259.51
RT4	30.65	375.59	78.44	375.59	24.36	375.59	50.38	375.59
RT5	103.86	1924.54	214.76	1924.54	58.79	1924.54	127.55	1924.54
HI5	3.68	7.85	3.33	7.85	2.11	7.85	4.71	7.85
RT6	41.34	438.89	45.82	438.89	26.58	438.89	54.37	438.89
RT7	76.29	1152.64	77.86	1152.64	49.89	1152.64	102.65	1152.64
RT8	117.01	1545.75	47.70	1545.75	49.09	1545.75	114.04	1545.75
HI8	14.92	12.56	1.00	12.56	2.94	12.56	5.58	12.56
RT9	117.20	649.66	99.77	649.66	33.64	649.66	75.27	649.66
HI9	29.12	7.66	8.33	7.66	2.72	7.66	5.57	7.66

Estimated means and variances of response time and highlight events variables for the final model are provided in Table 6. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. Based solely on response time, results suggest that class one progressed through the testlet faster than class two. On average there was a 30 second difference in the item response times between class one (fast responders) and class two (slow responders), such that fast responders spent 30 seconds less than slow responders on each item.

Figure 3. Stage One Testlet One Estimated Means of Process Data Variables (C2)

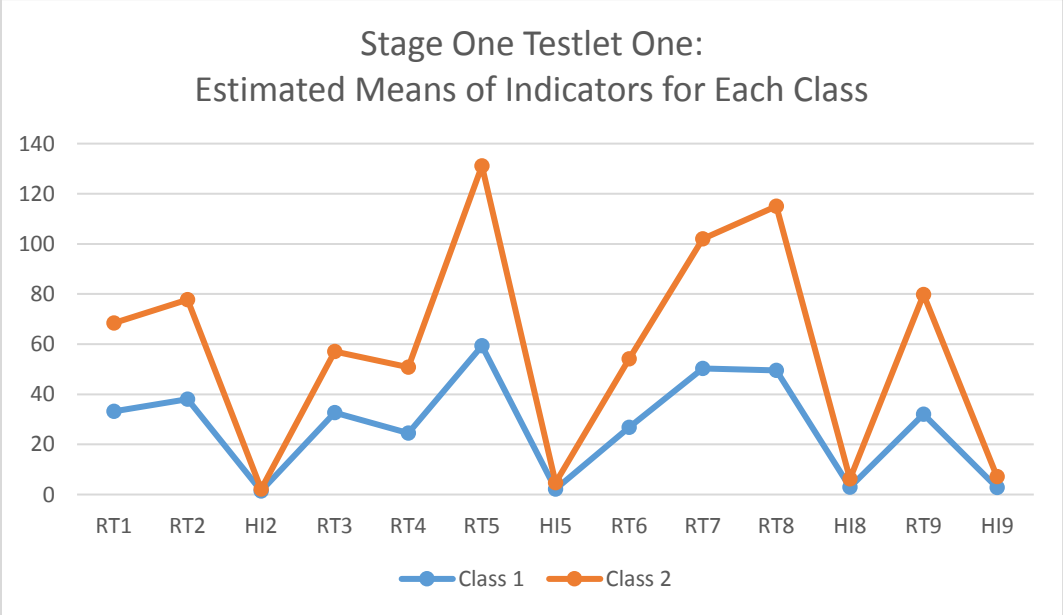

Table 6. Stage One Testlet One Estimated Means and Variances (C2)

	Class 1 (n = 677)		Class 2 (n = 143)	
Variable	Mean	Variance	Mean	Variance
RT1	33.14	855.80	68.38	855.80
RT2	38.03	1023.38	77.76	1023.38
HI2	1.44	1.45	2.15	1.45
RT3	32.63	492.59	57.12	492.59
RT4	24.50	381.00	50.74	381.00
RT5	59.30	1957.73	131.13	1957.73
HI5	2.13	7.87	4.71	7.87
RT6	26.81	442.24	54.14	442.24
RT7	50.31	1165.61	101.99	1165.61
RT8	49.59	1563.69	115.02	1563.69
HI8	2.95	13.59	6.25	13.59
RT9	32.04	676.15	79.72	676.15
HI9	2.77	14.09	7.20	14.09

The fast responders class contained 83% of the respondents while the slow responders class contained 17% of respondents. The largest difference between the two classes was found on the average time spent on item 5 and item 8. Results suggest that respondents approached the items differently in terms of response time but not highlight events.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.58, p = 0.44$). The test for race/ethnicity resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 1.02,\ p = 0.75$). A

three-way test for class, gender, and race/ethnicity was also examined to see if there was any significant interaction between the variables. The three-way chi square test was nonsignificant ($\chi^2(1) = 0.58, p = 0.44$).

Independent Samples T-Test

The average literacy score was calculated for each class as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy score between the classes (class one M = 269, SD = 44.73; class two M = 270, SD = 45.50; $t(818) = 0.25,\ p = 0.79$).

Overall, for stage one testlet one, fast responders spent half as much time as slow responders on items. However, the two classes were similar in terms of their highlight events, demographic makeup and literacy performance.

Stage One Testlet Two

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 7. Stage one testlet two contained 868 respondents. There were 403 (46%) males and 465 (54%) females in the testlet sample. By race/ethnicity, there were 546 (63%) white, 106 (12%) Hispanic, 147 (17%) black, and 69 (8%) other respondents included in the sample. Most respondents came from city 321 (37%) and suburban 252 (29%) locations. Forty-six percent of the sample was between the age of 16 and 34, while the other 54% of the sample ranged in age from 44 to 71 and above.

Table 7. Stage One Testlet Two Demographic Variables

Gender	n	%	Age	n	%
M	403	46%	16-19	102	12%
F	465	54%	20-24	100	12%
Location	n	%	25-29	88	10%
City	321	37%	30-34	103	12%
Suburb	252	29%	35-39	53	6%
Town	87	10%	40-44	61	7%
Rural	208	24%	45-49	79	9%
Race	n	%	50-54	66	8%
Hispanic	106	12%	55-59	64	7%
White	546	63%	60-65	56	6%
Black	147	17%	65-70	66	8%
Other	69	8%	71+	30	3%

Table 8 presents the descriptive statistics for all quantitative variables. On average, respondents spent close to a minute on the items and had less than four highlight events. Respondents also had an average score of 267 on the literacy section, which is a level two on the achievement level scale.

Table 8. Stage One Testlet Two Quantitative Variables

	RT1	RT2	RT3	RT4	HI4	RT5	RT6	HI6
Mean	54.16	33.90	38.69	74.41	2.58	67.63	65.75	3.62
SD	44.44	32.55	29.00	48.62	2.93	46.38	46.96	4.18
Min	2.04	2.61	2.17	2.00	0.00	1.99	3.66	0.00
Max	528.24	536.25	408.80	515.43	26.00	423.84	446.21	56.00
Skewness	3.24	7.11	4.35	1.99	3.04	2.58	2.56	4.96
Kurtosis	20.47	86.30	37.48	9.29	13.70	11.41	12.16	42.23

	RT7	HI7	RT8	HI8	RT9	HI9	AvgPV
Mean	42.96	3.33	83.19	2.32	84.70	2.91	266.95
SD	33.18	3.06	58.91	2.38	63.83	3.29	47.42
Min	2.60	0.00	2.56	0.00	2.60	0.00	95.07
Max	527.15	24.00	518.22	26.00	399.56	47.00	387.37
Skewness	4.90	2.71	2.27	3.88	1.46	5.38	-0.29
Kurtosis	55.24	10.66	9.33	26.62	2.50	54.92	-0.02

Final Model

Latent profile analysis for the data in stage one testlet two was conducted using a two-, three-, and four-class model. Fit statistics are provided in Table 9. Likelihood ratio tests were inconclusive and did not clearly favor one model over another. Although the four-class model has the lowest AIC and BIC, and the three-class model the highest entropy value, there was not enough information to distinguish classes from each other in these models. Therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 9. Stage One Testlet Two Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	101773.10	101978.04	0.88	0.44	< 0.01
3-Class Solution	101244.58	101521.02	0.93	0.22	< 0.01
4-Class Solution	100743.60	101091.53	0.91	0.06	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 10. Figure 4 plots the estimated means of the response time and highlight events variables. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. Based solely on response time, results suggest that class one progressed through the testlet faster than class two. On average there was almost a 50 second difference in the item response times between the fast responders and the slow responders, such that fast responders completed the testlet nearly twice as fast as slow responders.

Figure 4. Stage One Testlet Two Estimated Means of Process Data Variables

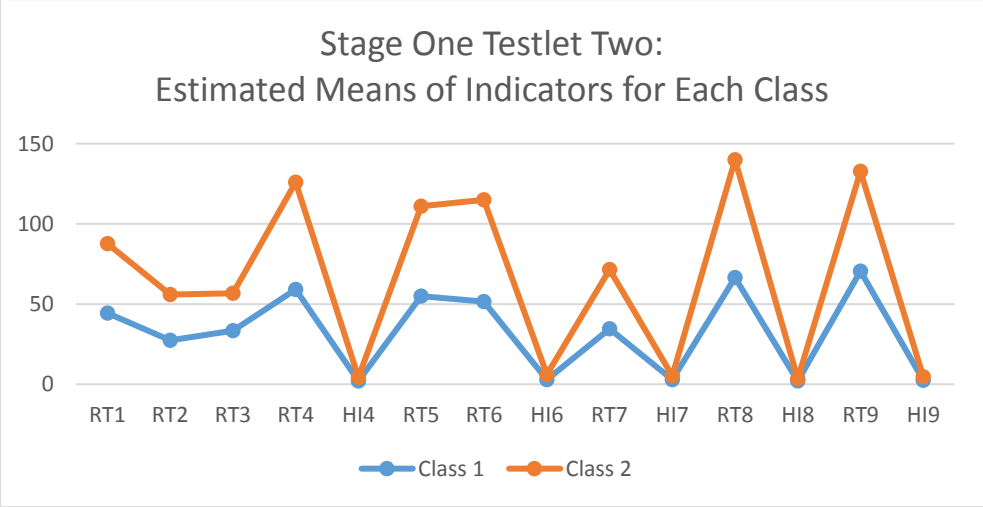

Table 10. Stage One Testlet Two Estimated Means and Variances

	Class 1 (n = 699)		Class 2 (n = 199)	
Variable	Mean	Variance	Mean	Variance
RT1	44.29	1642.52	87.62	1642.52
RT2	27.40	915.26	55.93	915.26
RT3	33.37	744.39	56.73	744.39
RT4	59.21	1577.43	125.96	1577.43
HI4	2.05	7.67	4.35	7.67

RT5	54.85	1595.14	110.95	1595.14
RT6	51.50	1514.68	114.95	1514.68
HI6	2.92	15.86	5.97	15.86
RT7	34.57	861.33	71.39	861.33
HI7	2.77	8.29	5.20	8.29
RT8	66.46	2516.95	139.92	2516.95
HI8	2.00	5.33	3.38	5.33
RT9	70.52	3387.69	132.78	3387.69
HI9	2.41	9.98	4.60	9.98

Seventy-seven percent of the respondents were fast responders while 23% of the respondents were slow responders. The largest difference between the two classes was found on the average time spent on item 4, item 5, item 6, item 8, and item 9. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time but not highlight events.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.76, p = 0.38$). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 2.85,\ p = 0.41$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy

score between the classes (class one M = 267, SD = 48.38; class two M = 264, SD = 44.07; $t(866) = 0.79,\ p = 0.42$).

Overall, for stage one testlet two, there was some difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, demographic makeup and literacy performance.

Stage One Testlet Three

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 11. Stage one testlet three contained 943 respondents. There were 437 (46%) males and 506 (54%) females in the testlet sample. By race/ethnicity, there were 567 (60%) white, 124 (13%) Hispanic, 182 (19%) black, and 70 (7%) other respondents included in the sample. Nearly 70% of respondents came from city and suburban locations. Forty-nine percent of the sample was between the age of 16 and 34, while the other 51% of the sample ranged in age from 44 to 71 and above.

Table 11. Stage One Testlet Three Demographic Variables

Gender	n	%	Age	n	%
M	437	46%	16-19	114	12%
F	506	54%	20-24	136	14%
Location	n	%	25-29	106	11%
City	385	41%	30-34	115	12%
Suburb	262	28%	35-39	76	8%
Town	94	10%	40-44	66	7%
Rural	202	21%	45-49	74	8%
Race	n	%	50-54	68	7%
Hispanic	124	13%	55-59	51	5%
White	567	60%	60-65	82	9%
Black	182	19%	65-70	35	4%
Other	70	7%	71+	20	2%

Table 12 presents the descriptive statistics for all quantitative variables. On average, respondents spent a little over a minute on the items and had less than three highlight events. Respondents also had an average score of 269 on the literacy section, which is a level two on the achievement level scale.

Table 12. Stage One Testlet Three Quantitative Variables

	RT1	RT2	HI2	RT3	HI3	RT4	HI4
Mean	53.41	104.32	3.12	95.88	3.48	87.02	1.52
SD	41.59	71.24	3.09	68.02	3.52	59.93	1.76
Min	2.33	2.84	0.00	2.56	0.00	2.29	0.00
Max	331.55	564.47	33.00	637.14	36.00	505.40	13.00
Skewness	2.28	1.74	2.80	1.93	3.53	1.87	2.68
Kurtosis	7.60	4.41	14.70	7.42	21.21	6.82	10.28

	RT5	RT6	HI6	RT7	HI7	RT8	HI8
Mean	39.58	74.79	1.49	67.95	1.74	69.56	1.50
SD	29.38	59.16	1.32	42.34	2.02	58.20	1.63
Min	2.07	4.77	0.00	2.18	0.00	2.15	0.00
Max	329.86	419.24	12.00	344.05	30.00	544.65	24.00
Skewness	3.17	2.09	3.35	2.00	6.67	2.84	6.60
Kurtosis	19.24	5.88	16.57	7.46	69.05	12.95	71.34

	RT9	HI9	AvgPV
Mean	55.70	1.58	268.90
SD	41.21	1.61	45.50
Min	2.13	0.00	126.45
Max	466.91	20.00	375.02
Skewness	2.66	5.52	-0.33
Kurtosis	14.90	50.36	-0.24

Final Model

Latent profile analysis for the data in stage one testlet three was conducted using a two-, three-, and four-class model. Fit statistics for each class are in Table 13. Likelihood ratio tests supported the two-class model, while AIC, BIC, entropy supported the four-class model. There was not enough separation between classes in the four-class model; therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 13. Stage One Testlet Three Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	116953.25	117190.85	0.88	< 0.01	< 0.01
3-Class Solution	116376.81	116696.85	0.92	0.37	< 0.01
4-Class Solution	115856.72	116259.19	0.94	0.53	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 14. Figure 5 plots the estimated means for the variables. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. On average there was almost a 60 second difference in the item response times between class one and class two, such that fast responders spent nearly 60 seconds less than slow responders on each item.

Figure 5. Stage One Testlet Three Estimated Means of Process Data Variables

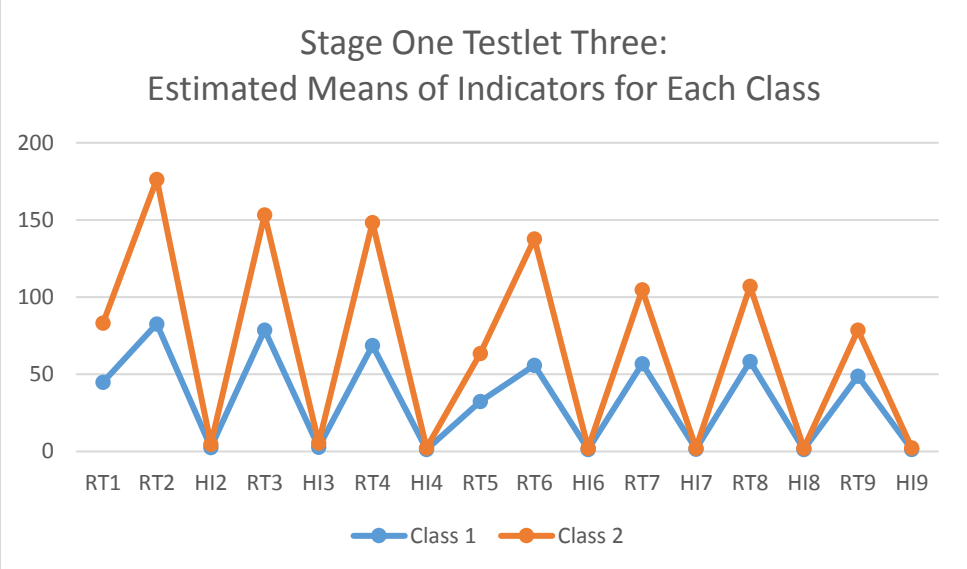

Table 14. Stage One Testlet Three Estimated Means and Variances

	Class 1 (n = 719)		Class 2 (n = 224)	
Variable	Mean	Variance	Mean	Variance
RT1	44.84	1462.58	83.14	1462.58
RT2	82.69	3512.25	176.33	3512.25
HI2	2.66	8.84	4.61	8.84
RT3	78.61	3630.36	153.35	3630.36
HI3	2.86	11.10	5.53	11.10
RT4	68.60	2459.36	148.31	2459.36
HI4	1.24	2.85	2.42	2.85
RT5	32.38	690.22	63.53	690.22
RT6	55.89	2306.89	137.73	2306.89
HI6	1.29	1.60	2.17	1.60
RT7	56.91	1385.34	104.69	1385.34
HI7	1.58	4.00	2.24	4.00
RT8	58.33	2963.53	106.96	2963.53
HI8	1.31	2.53	2.15	2.53
RT9	48.77	1537.15	78.73	1537.15
HI9	1.34	2.39	2.38	2.39

Class one contained 77% of the respondents while class two contained 23% of the respondents. The largest difference between the two classes were found on the average time spent on item 2, item 3, item 4, and item 6. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time but not highlight events.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.34, p =$

0.55). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 4.25,\ p = 0.23$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy score between the classes (class one M = 268, SD = 45.56; class two M = 269, SD = 45.38; $t(941) = -0.30,\ p = 0.76$).

Overall, for stage one testlet two, there was a minute difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, demographic makeup and literacy performance.

Stage Two Testlet One

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 15. Stage two testlet one contained 500 respondents. There were 224 (45%) males and 276 (55%) females in the testlet sample. By race/ethnicity, there were 305 (61%) white, 64 (13%) Hispanic, 95 (19%) black, and 36 (7%) other respondents included in the sample. Seventy-percent of respondents came from city and suburban locations. Forty-five percent of the sample was between the age of 16 and 34, while the other 55% of the sample ranged in age from 44 to 71 and above.

Table 15. Stage Two Testlet One Demographic Variables

Gender	n	%	Age	n	%
M	224	45%	16-19	55	11%
F	276	55%	20-24	74	15%
Location	n	%	25-29	46	9%
City	212	42%	30-34	49	10%
Suburb	141	28%	35-39	41	8%
Town	39	8%	40-44	36	7%
Rural	108	22%	45-49	41	8%
Race	n	%	50-54	39	8%
Hispanic	64	13%	55-59	40	8%
White	305	61%	60-65	42	8%
Black	95	19%	65-70	23	5%
Other	36	7%	71+	14	3%

Table 16 presents the descriptive statistics for all quantitative variables. On average, respondents spent close to a minute on the items and had less than three highlight events. Respondents also had an average score of 269 on the literacy section, which is a level two on the achievement level scale.

Table 16. Stage Two Testlet One Quantitative Variables

	RT1	HI1	RT2	HI2	RT3	HI3	RT4
Mean	27.21	1.16	57.49	3.26	83.46	3.00	40.01
SD	22.34	0.92	35.57	2.63	56.60	2.78	40.87
Min	2.66	0.00	2.13	0.00	2.20	0.00	1.96
Max	171.40	14.00	235.59	26.00	395.22	19.00	549.15
Skewness	3.20	8.58	1.18	2.46	1.55	2.02	5.51
Kurtosis	12.75	98.70	1.90	12.64	4.15	5.51	53.35

	RT5	RT6	HI6	RT7	HI7	RT8	RT9
Mean	28.34	51.88	2.55	56.89	1.45	43.28	33.54
SD	27.12	36.94	2.48	38.05	1.41	61.34	26.41
Min	2.24	2.05	0.00	3.48	0.00	1.53	1.57
Max	296.74	244.70	23.00	291.20	16.00	773.14	223.75
Skewness	5.44	1.68	3.14	1.76	4.64	6.75	2.47

Kurtosis	42.51	4.01	17.08	5.02	37.17	63.45	9.50

	HI9	RT10	HI10	RT11	HI11	AvgPV
Mean	1.53	64.56	1.23	60.52	2.03	269.31
SD	1.74	44.85	1.14	38.75	2.12	42.84
Min	0.00	9.46	0.00	9.55	0.00	132.60
Max	21.00	345.59	10.00	289.97	31.00	379.36
Skewness	4.72	1.88	3.41	2.09	6.22	-0.10
Kurtosis	37.32	5.94	18.88	6.89	72.18	-0.28

Final Model

Latent profile analysis for the data in stage two testlet one was conducted using a two-, three-, and four-class model. Fit statistics for each class are in Table 17. Likelihood ratio tests were inconclusive and did not clearly favor one model over another. Although the four-class model has the lowest AIC and BIC, and the highest entropy value, there was not enough information to distinguish classes from each other in this model. Therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 17. Stage Two Testlet One Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	70353.75	70598.19	0.91	0.37	< 0.01
3-Class Solution	69820.19	70148.93	0.94	0.59	< 0.01
4-Class Solution	69307.60	69720.63	0.95	0.21	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 18 and Figure 6 plots the estimated means of the variables. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. Class one spent less

time on all items in this testlet compared to class two. On average there was almost a 42 second difference in the item response times between class one and class two, such that fast responders spent nearly 42 seconds less than slow responders on each item.

Figure 6. Stage Two Testlet One Estimated Means of Process Data Variables

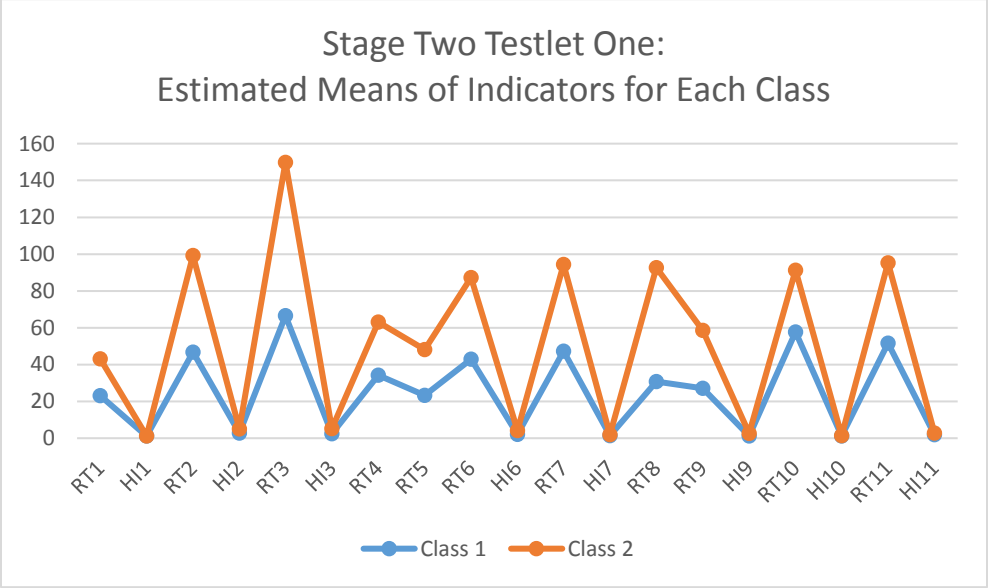

Table 18. Stage Two Testlet One Estimated Means and Variances

	Class 1 (n = 397)		Class 2 (n = 103)	
Variable	Mean	Variance	Mean	Variance
RT1	23.14	433.24	43.16	433.24
HI1	1.14	0.85	1.20	0.85
RT2	46.82	816.49	99.31	816.49
HI2	2.80	6.10	5.04	6.10
RT3	66.51	2072.01	149.88	2072.01
HI3	2.42	6.42	5.23	6.42
RT4	34.21	1530.66	63.11	1530.66
RT5	23.31	634.81	48.05	634.81
RT6	42.87	1043.02	87.22	1043.02
HI6	2.12	5.45	4.21	5.45
RT7	47.33	1087.45	94.33	1087.45
HI7	1.33	1.92	1.90	1.92

RT8	30.70	3135.24	92.59	3135.24
RT9	27.16	536.01	58.57	536.01
HI9	1.25	2.74	2.59	2.74
RT10	57.74	1824.74	91.31	1824.74
HI10	1.19	1.29	1.35	1.29
RT11	51.66	1190.73	95.27	1190.73
HI11	1.86	4.40	2.68	4.40

The fast responder class contained 80% of the respondents while the slow responder class contained 20% of the respondents. The largest difference between the two classes were found on the average time spent on item 3, and item 8. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time, although this separation is minimal.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.001, p = 0.74$). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 0.60,\ p = 0.89$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was a significant difference in the average literacy score between the classes (class one M = 272, SD = 42.63; class two M = 258, SD =

42.13; $t(498) = 2.84,\ p = 0.005, d = 0.32$). Although both classes had average literacy values on level two, fast responders scored 14 points higher than slow responders.

Overall, for stage two testlet one, there was a not a large difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, and demographic makeup.

Stage Two Testlet Two

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 19. Stage two testlet two contained 694 respondents. There were 318 (46%) males and 376 (54%) females in the testlet sample. By race/ethnicity, there were 422 (61%) white, 77 (11%) Hispanic, 77 (11%) black, and 52 (7%) other respondents included in the sample. Seventy-percent of respondents came from city and suburban locations. Half of the sample was between the age of 16 and 34, while the other half of the sample ranged in age from 44 to 71 and above.

Table 19. Stage Two Testlet Two Demographic Variables

Gender	n	%	Age	n	%
M	318	46%	16-19	68	10%
F	376	54%	20-24	94	14%
Location	n	%	25-29	81	12%
City	276	40%	30-34	96	14%
Suburb	187	27%	35-39	61	9%
Town	74	11%	40-44	45	6%
Rural	157	23%	45-49	43	6%
Race	n	%	50-54	55	8%
Hispanic	77	11%	55-59	43	6%
White	422	61%	60-65	52	7%
Black	143	21%	65-70	41	6%
Other	52	7%	71+	15	2%

Table 20 presents the descriptive statistics for all quantitative variables. On average, respondents spent close to a minute on the items and had less than three highlight events. Respondents also had an average score of 269 (level two) on the literacy section.

Table 20. Stage Two Testlet Two Quantitative Variables

	RT1	HI1	RT2	RT3	HI3	RT4	HI4
Mean	59.89	1.53	38.65	68.07	1.20	34.01	1.48
SD	42.06	1.43	37.86	46.17	1.15	26.04	1.21
Min	3.51	0.00	1.76	1.63	0.00	1.49	0.00
Max	372.44	20.00	346.82	364.92	16.00	244.13	10.00
Skewness	2.58	5.09	3.55	1.79	4.79	2.79	2.70
Kurtosis	10.36	47.19	19.14	5.23	45.62	13.27	10.62

	RT5	HI5	RT6	HI6	RT7	HI7	RT8
Mean	84.70	3.00	48.85	2.43	98.61	1.18	102.12
SD	52.45	2.76	37.58	2.42	74.60	1.00	84.11
Min	2.16	0.00	1.89	0.00	5.08	0.00	3.03
Max	472.03	23.00	406.69	26.00	630.40	10.00	618.75
Skewness	1.81	2.62	2.58	3.64	2.17	2.96	1.46
Kurtosis	8.51	10.88	14.50	21.71	9.50	17.23	3.39

	HI8	RT9	RT10	RT11	AvgPV
Mean	1.83	81.18	42.25	30.07	266.87
SD	2.22	50.06	44.58	23.77	48.15
Min	0.00	3.86	1.33	2.90	115.68
Max	19.00	348.55	656.76	221.50	395.99
Skewness	3.04	1.06	6.93	2.68	-0.42
Kurtosis	12.98	2.09	80.51	12.28	0.00

Final Model

Latent profile analysis for the data in stage two testlet one was conducted using a two-, three-, and four-class model. Fit criteria are provided in Table 21. Likelihood ratio tests were inconclusive and did not clearly favor one model over another. Although the four-class model has the lowest AIC and BIC, and the two- and three-class model had the highest entropy values, there was not enough information to distinguish classes from each other in the three and four-class model. Therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 21. Stage Two Testlet Two Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	96446.43	96696.26	0.90	0.41	< 0.01
3-Class Solution	95905.03	96241.18	0.90	0.71	< 0.01
4-Class Solution	95861.60	95566.31	0.85	0.67	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 22. Figure 7 plots the estimated means. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. On average there was almost a 50 second

difference in the item response times between class one and class two, such that fast responders spent nearly 50 seconds less than slow responders on each item.

Figure 7. Stage Two Testlet Two Estimated Means of Process Data Variables

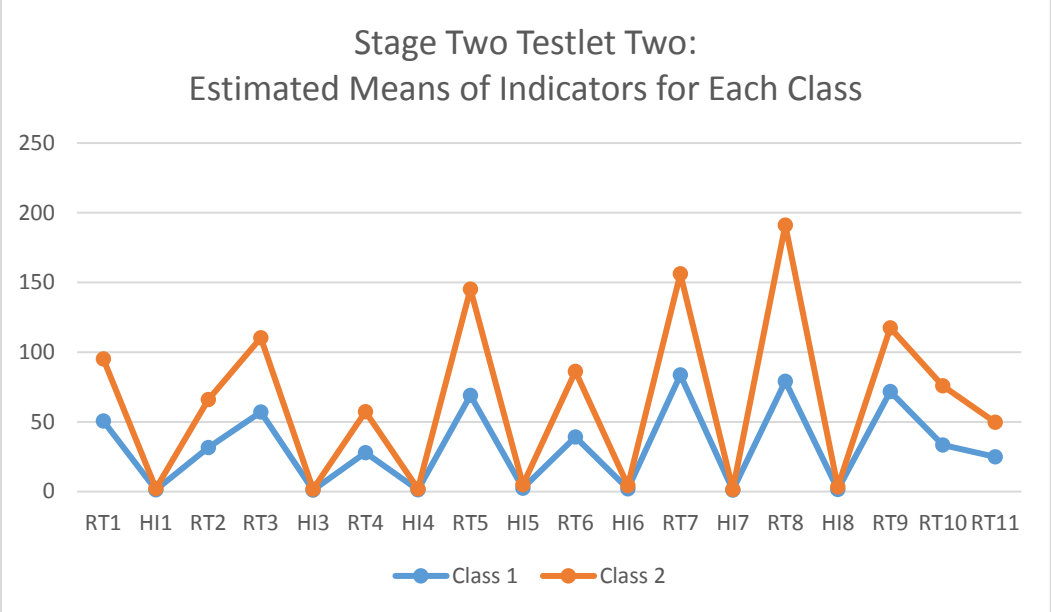

Table 22. Stage Two Testlet Two Estimated Means and Variances

	Class 1 (n = 554)		Class 2 (n = 140)	
Variable	Mean	Variance	Mean	Variance
RT1	50.70	1442.01	95.23	1442.01
HI1	1.35	1.91	2.20	1.91
RT2	31.53	1236.77	66.03	1236.77
RT3	57.08	1664.06	110.33	1664.06
HI3	1.07	1.25	1.71	1.25
RT4	27.96	536.38	57.28	536.38
HI4	1.35	1.40	1.97	1.40
RT5	68.99	1798.68	145.11	1798.68
HI5	2.42	6.58	4.98	6.58
RT6	39.13	1046.86	86.25	1046.86
HI6	1.93	4.90	4.34	4.90
RT7	83.61	4690.97	156.32	4690.97
HI7	1.06	0.94	1.60	0.94

RT8	78.98	5003.52	191.15	5003.52
HI8	1.42	4.30	3.38	4.30
RT9	71.76	2160.82	117.40	2160.82
RT10	33.51	1690.71	75.84	1690.71
RT11	24.94	463.32	49.77	463.32

Class one contained 80% of the respondents while class two contained 20% of the respondents. The largest difference between the two classes were found on the average time spent on item 5, and item 8. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time, although this separation is minimal.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.62, p = 0.43$). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 3.27,\ p = 0.35$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy score between the classes (class one M = 266, SD = 48.16; class two M = 268, SD = 48.22; $t(692) = 0.46,\ p = 0.64$).

Overall, for stage two testlet two, there was some difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, demographic makeup, and literacy performance.

Stage Two Testlet Three

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 23. Stage two testlet three contained 686 respondents. There were 307 (45%) males and 379 (55%) females in the testlet sample. By race/ethnicity, there were 429 (63%) white, 90 (13%) Hispanic, 121 (18%) black, and 46 (7%) other respondents included in the sample. Nearly 70% of respondents came from city and suburban locations. Half of the sample was between the age of 16 and 34, while the other half of the sample ranged in age from 44 to 71 and above.

Table 23. Stage Two Testlet Three Demographic Variables

Gender	n	%	Age	n	%
M	307	45%	16-19	91	13%
F	379	55%	20-24	85	12%
Location	n	%	25-29	86	13%
City	267	39%	30-34	85	12%
Suburb	197	29%	35-39	38	6%
Town	70	10%	40-44	46	7%
Rural	152	22%	45-49	59	9%
Race	n	%	50-54	44	6%
Hispanic	90	13%	55-59	44	6%
White	429	63%	60-65	52	8%
Black	121	18%	65-70	37	5%
Other	46	7%	71+	19	3%

Table 24 presents the descriptive statistics for all quantitative variables. On average, respondents spent a little more than a minute on the items and had less than two highlight events. Respondents also had an average score of 270 on the literacy section, which is a level two on the achievement level scale.

Table 24. Stage Two Testlet Three Quantitative Variables

	RT1	HI1	RT2	HI2	RT3	HI3	RT4
Mean	96.30	1.42	115.23	2.04	88.62	1.25	58.98
SD	70.03	2.15	82.39	3.59	54.64	1.15	39.92
Min	5.14	0.00	3.44	0.00	3.21	0.00	1.50
Max	688.45	41.00	430.70	56.00	384.10	15.00	316.44
Skewness	2.55	11.49	1.06	8.91	1.75	4.79	1.76
Kurtosis	12.42	184.15	1.16	108.34	4.36	41.49	5.07

	HI4	RT5	HI5	RT6	HI6	RT7	HI7
Mean	1.25	32.26	1.57	72.21	2.35	90.16	2.66
SD	1.04	29.70	1.70	50.17	2.12	58.37	2.04
Min	0.00	3.60	0.00	2.03	0.00	9.68	0.00
Max	13.00	477.35	26.00	367.01	24.00	387.97	19.00
Skewness	4.93	6.69	6.68	1.33	3.92	1.69	2.32
Kurtosis	38.42	82.48	72.08	3.06	26.11	3.94	10.75

	RT8	RT9	HI9	RT10	HI10	RT11	AvgPV
Mean	114.92	39.87	1.29	87.89	1.21	82.38	270.01
SD	75.03	25.24	1.02	65.82	0.98	57.97	48.60
Min	9.53	9.57	0.00	7.12	0.00	0.88	92.03
Max	591.55	179.96	8.00	477.09	9.00	466.12	384.06
Skewness	1.66	1.87	2.45	1.86	2.92	1.87	-0.29
Kurtosis	5.16	4.54	9.58	5.29	14.30	6.24	-0.12

Final Model

Latent profile analysis for the data in stage two testlet three was conducted using a two-, three-, and four-class model. Fit statistics for each class are in Table 25. Likelihood ratio tests were inconclusive and did not clearly favor one model over another. Although

the four-class model has the lowest AIC and BIC, and the highest entropy values, there was not enough information to distinguish classes from each other in the four-class model. Therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 25. Stage Two Testlet Three Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	103185.77	103462.15	0.83	0.56	< 0.01
3-Class Solution	102495.29	102866.82	0.88	0.62	< 0.01
4-Class Solution	101931.05	102397.73	0.90	0.46	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 26 and Figure 8 plots the estimated means. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. On average there was almost a 56 second difference in the item response times between class one and class two, such that fast responders spent nearly 56 seconds less than slow responders on each item.

Figure 8. Stage Two Testlet Three Estimated Means of Process Data Variables

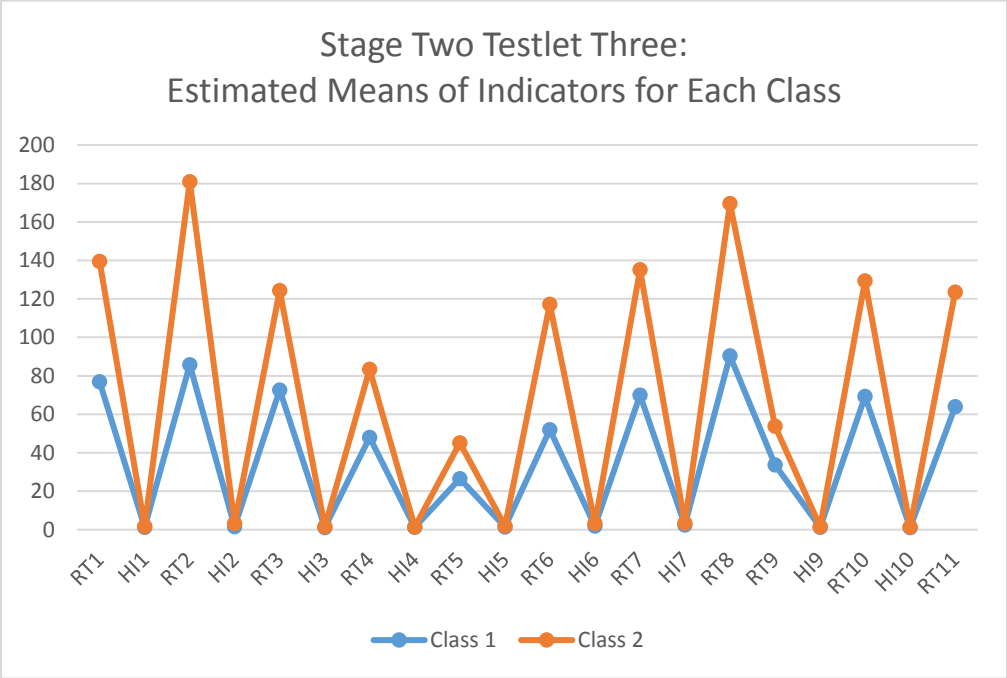

Table 26. Stage Two Testlet Three Estimated Means and Variances

	Class 1 (n = 471)		Class 2 (n = 215)	
Variable	Mean	Variance	Mean	Variance
RT1	76.90	4058.49	139.49	4058.49
HI1	1.26	4.55	1.76	4.55
RT2	85.67	4833.35	181.02	4833.35
HI2	1.54	12.34	3.14	12.34
RT3	72.52	2404.75	124.44	2404.75
HI3	1.09	1.27	1.59	1.27
RT4	47.99	1322.83	83.42	1322.83
HI4	1.17	1.08	1.40	1.08
RT5	26.48	806.77	45.11	806.77
HI5	1.41	2.83	1.92	2.83
RT6	52.00	1604.24	117.19	1604.24
HI6	1.95	4.11	3.25	4.11
RT7	69.94	2492.36	135.16	2492.36
HI7	2.31	3.89	3.42	3.89
RT8	90.35	4277.97	169.61	4277.97
RT9	33.64	549.72	53.72	549.72

HI9	1.20	1.01	1.48	1.01
RT10	69.29	3555.29	129.30	3555.29
HI10	1.11	0.95	1.41	0.95
RT11	63.89	2594.88	123.52	2594.88

The fast responders class contained 70% of the respondents while the slow responders class contained 30% of the respondents. The largest difference between the two classes were found on the average time spent on item 2, and item 8. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.28, p = 0.59$). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 4.31,\ p = 0.23$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy score between the classes (class one M = 268, SD = 46.73; class two M = 270, SD = 43.86; $t(684) = 0.58,\ p = 0.55$).

Overall, for stage two testlet three, there was a small difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, demographic makeup, and literacy performance.

Stage Two Testlet Four

Descriptive Statistics

Descriptive statistics for demographic variables are provided in Table 27. Stage two testlet four contained 756 respondents. There were 357 (47%) males and 399 (53%) females in the testlet sample. By race/ethnicity, there were 484 (64%) white, 91 (12%) Hispanic, 118 (16%) black, and 63 (8%) other respondents included in the sample. Sixty-three percent of respondents came from city and suburban locations. Nearly 50% of the sample was between the age of 16 and 34, while the other half of the sample ranged in age from 44 to 71 and above.

Table 27. Stage Two Testlet Four Demographic Variables

Gender	n	%	Age	n	%
M	357	47%	16-19	92	12%
F	399	53%	20-24	108	14%
Location	n	%	25-29	84	11%
City	265	35%	30-34	74	10%
Suburb	213	28%	35-39	51	7%
Town	84	11%	40-44	58	8%
Rural	194	26%	45-49	58	8%
Race	n	%	50-54	54	7%
Hispanic	91	12%	55-59	50	7%
White	484	64%	60-65	60	8%
Black	118	16%	65-70	44	6%
Other	63	8%	71+	23	3%

Table 28 presents the descriptive statistics for all quantitative variables. On average, respondents spent a about 60 seconds on the items and had less than two highlight events. Respondents also had an average score of 269 on the literacy section, which is a level two on the achievement level scale.

Table 28. Stage Two Testlet Four Quantitative Variables

	RT1	HI1	RT2	HI2	RT3	RT4	HI4
Mean	36.48	1.44	80.32	2.43	66.46	82.63	1.44
SD	28.91	1.14	53.56	2.07	43.01	61.26	1.21
Min	3.40	0.00	2.10	0.00	5.14	6.17	0.00
Max	205.18	11.00	409.54	19.00	438.04	761.41	10.00
Skewness	2.49	3.44	1.77	3.36	2.57	4.27	2.46
Kurtosis	8.27	15.50	6.22	17.62	13.29	38.80	8.73

Mean	RT5	HI5	RT6	RT7	HI7	RT8	HI8
Mean	98.63	2.81	113.24	42.64	1.35	107.05	2.84
SD	67.28	2.47	70.89	28.97	1.01	90.25	2.34
Min	9.30	0.00	9.54	9.70	0.00	7.34	0.00
Max	516.36	39.00	630.55	268.86	9.00	688.21	21.00
Skewness	2.13	5.70	1.86	2.68	2.49	2.32	2.63
Kurtosis	7.13	66.68	6.81	12.10	9.57	7.58	11.83

	RT9	HI9	RT10	RT11	AvgPV
Mean	104.98	1.86	82.76	56.83	268.79
SD	79.60	2.49	54.65	43.15	46.40
Min	0.43	0.00	0.88	0.38	95.07
Max	638.01	26.00	386.17	336.50	381.94
Skewness	2.06	5.34	1.77	1.90	-0.34
Kurtosis	7.93	38.77	4.87	5.97	0.06

Final Model

Latent profile analysis for the data in stage two testlet three was conducted using a two- and three-class model. Fit statistics are provided in Table 29. Likelihood ratio tests were inconclusive and did not clearly favor one model over another. Although the three-

class model has the lowest AIC and BIC, and the highest entropy values, there was not enough information to distinguish classes from each other in the three-class model. Therefore, based on interpretability of classes, the two-class model was chosen as the final model.

Table 29. Stage Two Testlet Four Fit Criteria

	AIC	BIC	Entropy	LMR-LRT	BLRT
2-Class Solution	108650.83	108940.38	0.89	0.09	< 0.01
3-Class Solution	108033.76	108376.24	0.92	NA	< 0.01
4-Class Solution	108495.99	108926.65	0.94	0.60	< 0.01

Estimated means and variances of response time and highlight events variables are provided in Table 30. Figure 9 plots the estimated means. On average class one spent less time on all items compared to class two; however; both classes on average had a similar number of highlight events on items. On average there was a 66 second difference in the item response times between class one and class two, such that fast responders spent nearly 66 seconds less than slow responders on each item.

Figure 9. Stage Two Testlet Four Estimated Means of Process Data Variables

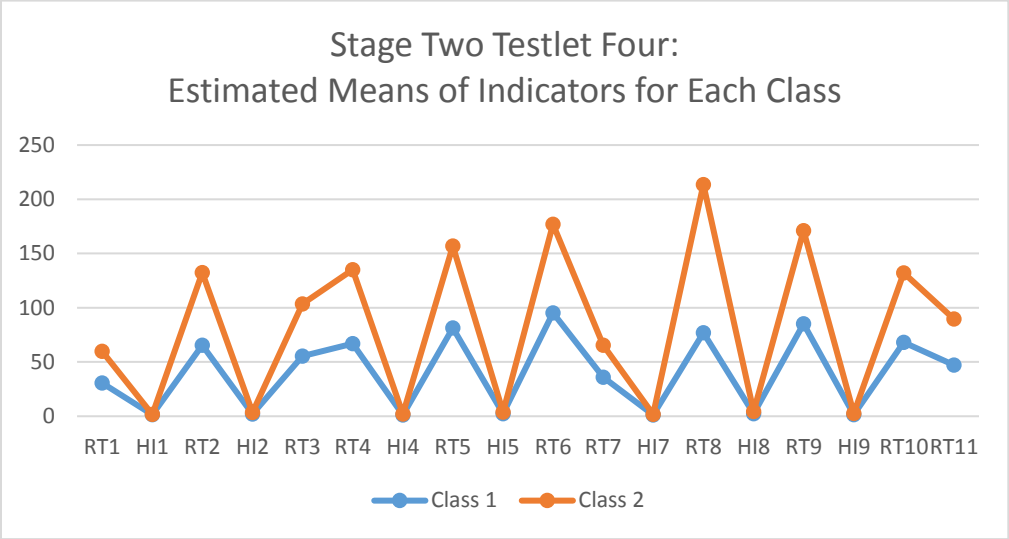

Table 30. Stage Two Testlet Four Estimated Means and Variances

	Class 1 (n = 576)		Class 2 (n = 180)	
Variable	Mean	Variance	Mean	Variance
RT1	30.60	710.32	59.95	710.32
HI1	1.37	1.28	1.64	1.28
RT2	65.40	2091.09	132.55	2091.09
HI2	2.14	3.99	3.40	3.99
RT3	55.55	1444.25	103.45	1444.25
RT4	67.05	2931.53	135.23	2931.53
HI4	1.27	1.37	2.00	1.37
RT5	81.51	3537.30	156.91	3537.30
HI5	2.48	5.74	3.90	5.74
RT6	95.35	3929.49	176.99	3929.49
RT7	36.07	694.75	65.55	694.75
HI7	1.24	0.99	1.85	0.99
RT8	77.01	5051.93	213.61	5051.93
HI8	2.36	4.70	4.42	4.70
RT9	85.20	4981.79	170.99	4981.79
HI9	1.58	5.97	2.74	5.97
RT10	68.23	2257.10	132.09	2257.10
RT11	47.17	1542.86	89.68	1542.86

The fast responders class contained 76% of the respondents while the slow responders class contained 24% of the respondents. The largest difference between the two classes were found on the average time spent on item 8, and item 9. The degree of separation between the two classes suggests that respondents approached the items differently in terms of response time.

Chi-Square Tests

Using class membership from the final two-class model, chi-square analyses were performed to assess the proportion of the demographic groups, gender and race/ethnicity in each class. The chi square test for gender resulted in a nonsignificant difference between males and females and their association with each class ($\chi^2(1) = 0.02, p = 0.86$). The test for race/ethnicity also resulted in a nonsignificant difference between race/ethnicity groups and their association with each class ($\chi^2(3) = 1.55,\ p = 0.67$).

Independent Samples T-Test

The average literacy score was calculated for each cluster as well. An independent samples t-test was calculated to examine if there was a significant difference between the average score for each class. There was not a significant difference in the average literacy score between the classes (class one M = 267, SD = 46.54; class two M = 271, SD = 45.90; $t(754) = -1.04,\ p = 0.29$).

Overall, for stage two testlet four, there was a difference between respondents' item response times between the two classes. However, the two classes were similar in terms of highlight events, demographic makeup, and literacy performance. A concise summary of the results section is provided in Table 31.

Table 31. Summary of Results

Statistics of Interest	Result
Final LPA Model	A two-class model was chosen as the best model for each testlet; interpretability and small sample size did not justify a model beyond two classes
Chi Square	No significant relationship between class assignment, gender, and race/ethnicity on any testlet
Two Sample T-Test	Except for stage two, testlet one, there was no significant difference in literacy achievement between classes

CHAPTER V

DISCUSSION

In the current educational landscape, researchers are not only interested in what students know but also how they learn. The investigation into how students learn has led researchers to use innovative methods to try to capture learning in real-time, using log-files or process data. Process data could show when students pause and for how long during a writing assessment, how students navigate through items (e.g. linearly or non-linearly), how often students look back or ahead to other items or parts of a test, or how much time students spend on items, among other things. The use of this type of data may provide insight into the behavioral processes of students (Lee & Haberman, 2016).

Patterns in these processes could be attributed to certain skills and learning strategies. Of interest in the literacy field is the use of reading and literacy strategies. Some of the reading strategies researchers have found evidence of include, skimming (Biedert et al., 2012), full-reading (Biedert et al., 2012), "search-and-destroy" (Huddleston & Lowe, 2014; Greaney, 2004; Heafner & Spooner, 2008), "look-backs" (Garner et al., 1984), and satisficing (Reader & Payne, 2007; Duggan & Payne, 2011), among others. Researchers want to know more about the intentional, cognitive, and behavioral aspects examinees use during reading assessments (Afflerbach, et al., 2008). Researchers have used eye-tracking technology, think-aloud protocols, survey responses and other self-report measures as data sources to examine how students use reading strategies during assessments. Although these data sources are useful, each have

their disadvantages. Eye-tracking technology is expensive and cumbersome to use with very large samples. Think-aloud protocols, survey-responses, and other self-report measures may lack reliability and are usually gathered after the fact (Prichard & Atkins, 2016). Process data from computer-based assessments has the advantage of collecting objective and unbiased data from any size sample, in real-time.

Latent class (LC) clustering is a latent variable model that attempts to uncover hidden, underlying groups from observed data (Oberski, 2016; Pastor et al., 2007). The use of LC clustering has increased in recent years due to its ability to handle large amounts of data (Fraley & Raftery, 1998), like process data. This study proposed a new design that used available process data collected from the Program for the International Assessment of Adult Competencies (PIAAC) computer-based assessment and LC clustering as way to identify patterns of behaviors potentially related to reading strategies.

The current study sought to ask two major research questions. The first question was, *can latent class (LC) analysis uncover latent classes that correspond to specific reading strategies?* In all testlets, a two-class model was chosen as the final model because there was not enough descriptive or statistical separation between classes to justify a three- or four-class model. In addition, the presence of a third or fourth class resulted in a small sample of less than five percent in each class. Apparent in all testlets was a class that progressed fast through each testlet (fast responders) and a class that progressed slower through each testlet (slow responders), based only on item response time. There was no difference in the use of item highlight events between classes. With

only the presence of two classes, it is difficult to solidly ascribe the patterns in each class to different reading strategies.

In a study by Biedert et al. (2012), researchers found evidence of two reading strategies: full-reading and skim-reading. Full-reading was characterized by a longer time duration while skim-reading was characterized by a shorter time duration. Using the results from Biedert et al. (2012) the two classes identified in the current testlet results might be classified as full-reading and skim-reading. However, in some testlet results there was not enough separation between the classes to describe them as strategies. For example, the two classes in stage one testlet one were only separated by an average of 30 seconds, in terms of item response time. The range in item response time difference between the two classes was 47 seconds. In contrast, the two classes in stage two testlet four were separated by an average of 66 seconds, in terms of item response time. The range in item response time difference between the two classes was 56 seconds.

It is also possible that the resulting two classes (fast responders and slow responders) are categories of a continuous latent variable, speed. If a factor analysis had been performed, it is possible that a single latent factor, possibly speed, could have emerged. Instead, with the given information, the latent profile analysis dichotomized speed, resulting in a fast responders class and a slow responders class.

The second major research question was, *do uncovered latent classes have a relationship with any prominent observed variables?* In all cases, there was no significant association between gender or race/ethnicity and class assignment. Therefore, males and females, and examinees of difference race/ethnic groups were found to be independent of class assignment. In all except one case, the average literacy score between the two

classes was nearly the same, whether examinees were fast responders or slow responders. In the one case where there was a significant difference in average literacy score, the fast responders scored 14 points higher (d = 0.32) than the slow responders. Given no difference in other testlets, this result should not be over interpreted, as both classes were still at a level two on the achievement scale.

Greaney (2004) and Heafner and Spooner (2008) found that examinees commonly used skimming like strategies but that these strategies were used unproductively. Results of the current study confirm the common use of skimming like strategies, such that the majority of examinees in each testlet belonged to the fast responders class. However, the result of unproductive use was not found in the current study as both classes had near similar literacy scores, regardless of using more or less time on each testlet.

The only variable that contributed to the description of each class was item response time. It is interesting to note that the degree of separation between classes increased with each testlet on stage one and stage two. The average item RT between classes was 30 seconds on stage one testlet one, 50 seconds on stage one testlet two, and 58 seconds on stage one testlet three. The average item RT between classes was 42 seconds on stage two testlet one, 52 seconds on stage two testlet two, 56 seconds on stage two testlet three, and 66 seconds on stage two testlet four. As testlet difficulty increased within each stage, so did the separation between classes in each testlet.

There was limited evidence in the current study for the use of reading strategies used during the PIAAC assessment, except for maybe the use of full-reading and skim-reading. Beidert et al. (2012) noted that full-reading was characterized by longer response time while skim-reading was characterized by shorter response time. However, there are

two potential implications of the current study. First, the fast responders group was no more able than the slow responders group so potentially speeding through an assessment could produce a speededness effect that is unfair (Lu & Sireci, 2007). Also, given that the PIAAC assessment is not timed, there is no consequence for taking more time and progressing slowly. Second, the consistency of the patterns shown from the fast responder and slow responder groups across all testlets and both stages suggest that the testlets are well constructed and there were no differential effects of individual items, testlets, or stages, that produced different patterns in the groups.

Aside from the proportion of people in each class and the speed at which respondents progressed through testlets for each class, the classes were found to be almost uniform and not have any significant relationship with gender, race/ethnicity, or difference in literacy score.

Limitations and Future Directions

While the current study does contribute to the field by proposing a new way of examining potential reading strategies, it is not without its limitations. The first limitation is that the PIAAC is a low-stakes assessment. One of the biggest problems facing low-stakes assessments is low test-taking engagement and motivation. Low-stakes assessments usually carry no consequence for performance for examinees, therefore, examinees may not give their best effort (Wise & DeMars, 2005). Taking into account this information, in the current study, examinees possibly did not engage in any behaviors that would reflect potential reading strategies because they were not engaged or motivated enough to do so. Low engagement and motivation may also lead to invalid scores as examinee performance can be underestimated on low-stakes assessments (Wise

& DeMars, 2005). This reasoning may also help to explain why the average literacy achievement level for each class in every testlet was a level two and not higher or show much variation.

The second limitation is that item response time was the only variable that could provide some distinction between classes in every model. The number of highlight events per item was not useful in detecting pattern differences between classes. Simply, there weren't enough useful process data variables used in the models and models are only as good as their indicators. Part of the problem is that there weren't many process data variables to choose from in the PIAAC data set. The LogDataAnalyzer tool (OECD, 2017) facilitates data processing and data extraction using log files in raw .xml format. However, limited predefined variables can be generated using the LogDataAnalyzer. Only two out of 19 variables were thought to be useful for assessing patterns from the PIAAC literacy section.

Another part of the problem is that there has not been much research done on process data in general but especially not on the types of variables extracted from process data and the utility of those variables. The use of process data is still new and researchers are thinking about the best ways to use process data meaningfully. Hobert et al., (2013) stated that there is a lack of theory for parsing and aggregating process data in a meaningful way. If there were a better understanding of process data and its extraction, other latent classes may have emerged. Still, response time is one of most used pieces of process data and has become one of the main focuses of process data research (van der Linden, 2009; Lee & Haberman, 2016; Wise & Kong, 2005; Molenaar, 2015; Kong, Wise, & Bhola, 2007).

The third limitation is that fit criteria measures were not very useful in helping choose final models. Ultimately, interpretability was the only criteria used in selecting final class models. In some testlet results, the degree of separation between classes was small; therefore, the fit criteria poorly selected the correct number of classes. In all, the limitations can be improved upon by conducting more research on the aforementioned topics.

The results and limitations of the current study warrant a better design in future research. First, focusing on the investigation of reading strategies in high stakes assessments, where examinees are more likely to be engaged or motivated to try their best, might lead to different results. Second, there is much more research that can be done with process data. Process data is a topic area that has greatly benefitted from the research that has already been produced and researchers are looking to push the area even further. The identification of new process data variables with explanatory power, in conjunction with more established variables like response time, could lead to more impactful outcomes.

Third, future research may benefit from setting a threshold for a minimum time to distinguish between examinees who use rapid-guessing and skimming (Goldhammer, Martens, Lüdtke, 2017). By setting a threshold to identify rapid-guessers, those who are identified as full-readers will be more distinguishable from both groups. Also, setting a time threshold would allow for response time to be better used as an indicator of a potential speed latent variable (Goldhammer & Entink, 2011).

Fourth, in the context of the current study once a base model is established, researchers may add covariates to the model. Item level, student level, or even test level

characteristics could be added to LC clustering models as covariates to better assess potential reading strategies. Last, if the current study is used with an adolescent sample, a longitudinal trajectory of potential reading strategies over time could contribute to research. Research could focus on estimating a turning point or intervention that would change the developmental trajectory of the use of reading strategies.

Conclusion

Like item and test scores, process data can be used to draw inferences about examinees and what they know (Goldhammer & Zehner, 2017). Descriptive performance profiles can be constructed for groups of students, that can provide information above and beyond test scores. Using process data to supplement test scores can help researchers determine how students are interacting with tests and if different student processes help or hurt those students' overall performance. Nuanced and rigorous models would contribute to the pursuit of understanding student knowledge and behaviors during assessments. The use of this type of data may provide insight into the behavioral processes of students. Specifically, incorporating reading strategy information with student level data could build better descriptive profiles of students to show strengths, weaknesses, or even target inventions. Using this information in conjunction with background questionnaires that asses student, teacher, and school characteristics, might uncover what characteristics are associated with different reading strategies.

REFERENCES

Adams, M. A., Sallis, J. F., Kerr, J., Conway, T. L., Saelens, B. E., Frank, L. D., ... & Cain, K. L. (2011). Neighborhood environment profiles related to physical activity and weight status: A latent profile analysis. *Preventive Medicine*, *52*(5), 326-331.

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364-373.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & B.F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Academiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal*, *75*(4), 460-472.

Andruff, H., Carraro, N., Thompson, A., & Gaudreau, P. (2009). Latent class growth modelling: A tutorial. *Tutorials in Quantitative Methods for Psychology*, *5*(1), 11-24.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., & Rathouz, P.J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association, 92*(440), 1375-1386.

Banfield, J. D., & Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

Berry, J. W. (1997). Immigration, acculturation, and adaptation. *Applied Psychology: An International Review, 46,* 5–34.

Biedert, R., Hees, J., Dengel, A., & Buscher, G. (2012). A robust realtime reading-skimming classifier. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 123-130). ACM.

Birkerts, S. (1994). *The gutenberg elegies: The fate of reading in an electronic age*. Boston, MA: Faber and Faber.

Boscardin, C. K. (2012). Profiling students for remediation using latent class analysis. *Advances in Health Science Education, 17,* 55-63.

Boulet, J. R., De Champlian, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OCSEs and standardized patient examinations. *Medical Teacher, 25*(3), 245–249.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370.

Brasseur-Hock, I.F., Hock, M.F., Kieffer, M.J., Biancarosa, G., & Deshler, D.D. (2011). Adolescent struggling readers in urban schools: Results of a latent class analysis. *Learning and Individual Differences, 21,* 438-452.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195-212.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson & E. F. Borgotta (Eds.), *Factor Analysis and Measurement in Sociological Research,* (pp. 215–246). Beverly Hills, CA: Sage.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Hoboken, NJ: John Wiley & Sons.

Dayton, C. M. (1991). Educational applications of latent class analysis. *Measurement & Evaluation in Counseling and Development, 24*(3), 131-142.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83*(401), 173-178.

Denton, C. A., Wolters, C. A., York, M. J., Swanson, E., Kulesz, P. A., & Francis, D. J. (2015). Adolescents' use of reading comprehension strategies: Differences related to reading proficiency, grade level, and gender. *Learning and Individual Differences, 37*, 81-95.

Duggan, G. B., & Payne, S. J. (2011). Skim reading by satisficing: Evidence from eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1141-1150.

Dutt, A., Ismail, M.A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.

Ercikan, K. (2018, February). *A Test Can Do That?* Retrieved from the Educational Testing Service website: http://news.ets.org/stories/a-test-can-do-that/

Finch, W. H., & Bronk, K. C. (2011). Conducting confirmatory latent class analysis using mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(1), 132-151.

Fraley, C., & Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578-588.

Garner, R., Hare, V.C., Alexander, P., Haynes, J., & Winograd, P. (1984). Inducing use of a text lookback strategy among unsuccessful readers. *American Educational Research Journal, 21*(4), 789-798.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 24, 229–252.

Gil, L., Martinez, T., & Vidal-Abarca, E. (2015). Online assessment of strategic literacy skills. *Computers & Education, 82*, 50-59.

Goldhammer, F., & Entink, R. H. K. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence, 39*(2-3), 108-119.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory irt modelling approach considering person and item characteristics. *Large-scale Assessments in Education, 5*(1), 18-42.

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3-4), 128-132.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231.

Greaney, K. (2004). Factors affecting pat reading comprehension performance: A retrospective analysis of some year 4 6 data. *New Zealand Journal of Educational Studies, 39*(1), 3-21.

Grunschel, C., Patrzek, J., & Fries, S. (2013). Exploring different types of academic delayers: A latent profile analysis. *Learning and Individual Differences, 23*, 225-233.

Guo, H., Rios, J. A., Haberman S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173-183. doi:10.1080/08957347.2016.1171766

Haberman, S. J. (1979). *Analysis of qualitative data, vol. 2: New developments*. New York, NY: Academic Press.

Hare., V. C. (1981). Readers' problem identification and problem-solving strategies for high- and low-knowledge articles. *Journal of Reading Behavior, 13*(4), 359-365.

Heafner, T. L., & Spooner, M. (2008). Promoting learning in a professional development school: Helping students "get over the mountain". In I. N. Guadarrama, J. M. Ramsey, & J. L. Nath (Eds.), *University and school connections: Research studies in professional development schools* (pp. 117– 150). Charlotte, NC: Information Age.

Hobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R.S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo and A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 1-13). Springer Press.

Huddleston, A. P., & Lowe, T. N. (2014). “I skim and find the answers”: Addressing search and destroy in reading. *The Reading Teachers, 68*(1), 71-79.

Hyönä, J., Lorch, R. F., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*(1), 44-55.

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Jorgensen, M., & Hunt, L. A. (1996). Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS*, *96*, 375-384.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth modeling and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302-317.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. doi:10.1177/0013164406294779

Lawrence, C.J., & Krzanowski, W.J. (1996). Mixture separation for mixed-mode data. *Statistics and Computing, 6*, 85-92.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, 362-412.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

Lee, Y-H., & Haberman, S.J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing, 16*(3), 240-267.

Li, L., Tseng, S., & Chen, G. (2016). Effect of hypertext highlighting on browsing, reading, and navigational performance. *Computers in Human Behavior, 54*, 318-325.

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation, 61*(6), 700-12.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29-37.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1*(14), 281-297.

Maddox, B. (2017). Talk and gesture as process data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3-4), 113-127.

Magidson, J., & Vermunt, J. K. (2002a). Latent class modeling as a probabilistic extension of k-means clustering. *Quirk's Marketing Research Review*, 20, 77-80.

Magidson, J., & Vermunt, J. K. (2002b). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research, 20*(1), 36-43.

Mahdavi, A., & Azimi, S. (2012). The effects of cognitive strategies i.e. note-making and underlining on Iranian efl learners' reading comprehension. *International Journal of Applied Linguistics & English Literature, 1*(6), 1-6.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. *Journal of the Royal Statistical Society, 36*(3), 318-324.

McLachlan, G. J., & Basford, K. E. (1988) *Mixture models: Inference and applications to clustering*. New York City, NY: Marcel Dekker.

McLachlan, G. J., & Chang, S. U. (2004). Mixture modelling for cluster analysis. *Statistical Methods in Medical Research, 13*(5), 347-361.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York City, NY: John Wiley & Sons, Inc.

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 177–181. doi:10.1080/15366367.2015.1105073

Muthén, B. O. (2000). Methodological issues in random coefficient growth modeling using a latent variable framework: Applications to the development of heavy drinking. In J. Rose, L. Chassin, C. Presson, & J. Sherman (Eds.), *Multivariate applications in substance use research* (pp. 113–140). Hillsdale, NJ: Erlbaum.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *New Developments and Techniques in Structural Equation Modeling* (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Muthén, L. K. and Muthén, B. O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1999). Re: Does Mplus impute values for those that are missing [Online comment]. Retrieved from http://www.statmodel.com/discussion/messages/22/24.html?1534981458

Muthén, L. K., & Muthén, B. O. (2001). Re: Latent profile analysis [Online comment]. Retrieved from http://www.statmodel.com/discussion/messages/13/115.html

Muthén, L. K., & Muthén, B. O. (2010). Re: Skewed variables for latent class/profile analysis [Online comment]. Retrieved from http://www.statmodel.com/discussion/messages/13/5439.html?1520375434

Muthén, L. K., & Muthén, B. O. (2017). Re: How to handle missing data in Latent Transition Analysis [Online comment]. Retrieved from http://www.statmodel.com/discussion/messages/13/23896.html?1524606352

National Center for Education Statistics. (2018, August). *Program for the International Assessment of Adult Competencies (PIAAC)*. Retrieved from the National Center for Education Statistics website: https://nces.ed.gov/surveys/piaac/

OECD (2017). Programme for the International Assessment of Adult Competencies (PIAAC), log files. GESIS Data Archive, Cologne. ZA6712 Data File Version 2.0.0, doi: 10.4232/1.12955

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI: A modern look at data analysis for HCI research* (pp. 275-287). Cham, Switzerland: Springer.

Pastor, D. A., Barron, K. E., Miller, B. J., Davis, S. L. (2007). A latent profile analysis of college student's achievement goal orientation. *Contemporary Educational Psychology, 32,* 8-47.

Pernice, K. (2017, November). *F-shaped Patter of Reading on the Web: Misunderstood, but Still Relevant (Even on Mobile)*. Retrieved from the Nielsen Norman Group website: https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

Prichard, C., & Atkins, A. (2016). Evaluating L2 readers' previewing strategies using eye-tracking. *The Reading Matrix: An International Online Journal, 16*(2), 110-130.

Reader, W. R., & Payne, S. J. (2007). Allocating time across multiple texts: Sampling and satisficing. *Human Computer Interaction, 22,* 263-298.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis,* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc. ISBN-13: 978-0-470-17896-6.

Sainsbury, M., & Benton, T. (2011). Designing a formative e-assessment: Latent class analysis of early reading skills. *British Journal of Educational Testing, 42*(3), 500-514.

Schwartz, S. J., & Zamboanga, B. L. (2008). Testing berry's model of acculturation: A confirmatory latent class approach. *Cultural Diversity and Ethnic Minority Psychology, 14(*4), 275-285.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Scott, A. J., & Symons, M. J. (1971) Clustering methods based on likelihood ratio criteria. *Biometrics*, 27, 387-397.

Silver, V. L., & Kreiner, D. S. (1997). The effects of pre-existing inappropriate highlighting on reading comprehension. *Reading Research and Instruction, 36*(3), 217-223.

Tein, J-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 640-657.

Templin, J. (2006). Mixture and Latent Class Analysis [PowerPoint slides]. Retrieved from jonathantemplin.com

van der Heijden, P. G. M., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the em algorithm. *Journal of Educational and Behavioral Statistics, 21*(3), 215-229.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272.

van der Schoot, M., Vasbinder, A. L., Horsley, T. M., & van Lieshout, E. (2008). The role of two reading strategies in text comprehension: An eye fixation study in primary school children. *Journal of Research in Reading, 31*(2), 203-223.

Vaughn, M.G., DeLisi, M., Beaver, K.M., & Howard, M.O. (2008). Toward a quantitative typology of burglars: A latent profile analysis of career offenders. *Journal of Forensic Sciences, 53*(6), 1387-1392.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450-469.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (pp. 1-10). Thousand Oaks, CA: SAGE Publications.

Wise, S., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, *5*(3), 329-350.

Wu, J. (2014). Gender differences in online reading engagement, metacognitive strategies, navigation skills and reading literacy. *Journal of Computer Assisted Learning*, *30*, 252-271.

Zhu, M., Shu, Z., & von Davier, A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement, 53*(2), 190-211.