Revealing Test Answer Behavior Patterns Through Quartile Analysis

LeAnne J. Schmidt, Central Michigan University, United States
Kathryn Dirkin, Central Michigan University, United States

The IAFOR Conference on Educational Research & Innovation 2022
Official Conference Proceedings

Abstract

Characterizing a dataset by the mean value homogenizes the data to lose the integrity of the highs and lows, however, a quartile analysis quantifies the tendencies of both high- and low-performing participants for comparison. This study analyzed the grammar assessment responses of 8th grade students to determine patterns of response between the lowest and highest quartile. Using Peng's Learning Portrait Model, each assessment cell was coded to show the accuracy of prior and subsequent answers. Analysis of these codes revealed that learners in the lowest quartile were significantly likely to respond inconsistently (variable accuracy, such as correct-incorrect-correct) and that learners in the highest quartile were significantly likely to respond consistently, whether correct or incorrect. Further, the baseline score increased over the course of seven months by 25% on unrelated content, suggesting that familiarity with the application software can account for that much of a student's assessment score. Future explorations on the dynamics of online assessment and the persistence of students in resolving inaccuracies on digital assessments are encouraged.

Keywords: Assessment, Quartile Analysis, Accuracy, Middle School, Grammar Assessment, Application Effects, Digital Assessment, Consistency, Inconsistence, Response Patterns

Introduction

Analysis of cohort data can be beneficial for generalizing responses of the overall group, however, unique features of subgroups are homogenized in the process. Adjusting the analysis to group the cohort by quartiles can reveal variations between the highest and lowest quartile, which would be entirely lost by averaging. In short, the midpoint between success and failure is moderate success, but the midpoint provides little valuable within-group information. Quartile analysis, however, establishes a profile for success at the top quartile and for poor performance at the lowest quartile. To educators who strive to guide learners on the path of success, an analysis of the variations between the higher and lower performances can help to identify key patterns which are common to low-performing students who differ from those of more successful students. Interruption of these problematic behavior patterns in responses can lead students toward growth. Training in methods which emulate the behavior patterns of success equip struggling learners to perform better.

In classes which engage in routine assessment in a certain style, the effect of quartile-based analysis and guidance can yield marked results. For learners at the lowest level, facing the greatest challenge, mapping a pathway to success is nothing short of shedding light to escape from a dark tunnel or providing a ladder to someone stranded at the bottom of a deep hole. Until teachers know what is unique to the situation of these learners in comparison to those at the top of the class, there can be no map, no light, and no ladder.

Literature Review

Importance of prompt, direct, high-quality feedback

Immediate, explicit, and well-grounded feedback is necessary for academic advancement. Study after study echoes the imperative that students need prompt, direct, high-quality feedback on their work to optimize improvement (Butler & Nisan, 1986; Kluger & DeNisi, 1996; Hattie & Timperley, 2007; Kulik & Kulik, 1988; Yang et al., 2014). These qualities are present in automated feedback provided by intelligent tutoring systems in a smart learning environment (VanLehn, 2011; Spector, 2016; Paassen, Mokhel, & Hammer, 2016; Slof et al., 2013; Koedinger et al., 2013). These responses to the answers that learners provide in online assessments are informative for students, but it is the overview and analysis of these assessments which inform teachers. Key questions following an assessment on which teachers need detailed feedback to meet the needs of the students and guide them effectively include: Who fell below the mean? Who scored well? Did students improve over the previous assessment? However, for a teacher to get answers to these questions, calculations are needed. Few teachers have or choose to expend the available time tabulating comparative data. A score can provide some information, but behavioral patterns, especially when studied longitudinally, can provide insight to their causes and other influential factors.

Value of AI in providing prompt response

The obstacle to timely feedback lies in ever-present barriers to classroom teaching. Though over 50 years old, a study by Claye (1968) defines these barriers in rank order: overcrowded classrooms, too much clerical work, pressure from supervisors and administrators, lack of instructional materials and supplies, too many interruptions, rigid curriculum, lack of freedom, rigid time schedule, and too many non-teaching duties. The same factors are shared in more recent studies as well (Gallo et al., 2006; Bresciani, 2011). Faced with these impediments, teachers are not able to provide immediate, accurate, and effective feedback to all students for all assignments. However, advances in technology and accessibility for students to computer-based or online resources opened new opportunities through the use of artificial intelligence (AI) and intelligent tutoring systems.

Computer-aided instruction (CAI) and virtual reality (VR) enhance the impacts of classroom teaching through new opportunities. Traditional methods pose limits, while VR bridges the gap and makes learning personal and interactive, offering skill training and authentic environments (Xie, 2018, p. 76). Extensions of CAI produced adaptive learning systems which offered personalization of experience based on the unique responses of each learner. Programming branching scenarios and levels of response allowed programs to meet the learner at any stage of development and work toward advances (Peng et al., 2019). This adaptive technology provided the opportunity to respond to the individual based on accumulated data, but also provides a wealth of data about comprehension and logical pathways supporting the *learning portrait model* (Peng et al., 2019).

Learning Portrait Model

The *Learning Portrait Model* described a sequence of learning cells which depict learning processes (Peng et al., 2019). Each cell includes a learning act and represents a duration necessary to experience that act (Peng et al., 2019). A sequence of cells represents a learning pattern (Peng et al., 2019). As the duration of the cell increases, the level of engagement also increases (Peng et al., 2019). An analysis of the cells which comprise an assignment, the duration of each, and the relative duration of each cell in a sequence can provide insight on the task itself. Further, a comparison of the duration of a cell (learning act) for different students can lead to valuable analysis in the differential factors at work, as well as which patterns characterize which profile of student, if there are commonalities. Each cell consists of a micro-learning situation including content, activities, and effects (Peng et al., 2019). The occurrences or *effects* in a cell can present different experiences of learners in the same learning setting, with the same content and assignment. Granular data is necessary to explore the variations within and among cells in a learning sequence. Such granular data includes each attempt within an assignment and a characterization of the learning activity for each experience. At this level, patterns of learning cell sequences may be recognizable and even predictive of learner behaviors.

Data-Informed, Personalized Formative Assessment

Within the architecture of a learning environment, several factors are necessary, whether they are traditional or the product of technology: context awareness, adaptive support, adaptive interface, adaptive content, personalized support, tracking and updating learner progress, and an inference or recommendation engine (Hwang, 2014). While these are well within the capabilities of educators, they are time-consuming and labor intensive on a one-to-one scale, but a virtual impossibility on a large scale, so some elements are sacrificed. With the automated options available, these sacrifices are no longer necessary. Technology resources are capable of providing many answers, but some are better suited to teacher expertise and relationship-building with learners. Partnership between technology and the educator can be valuable (Spector et al., 2016). In fact, "lack of teacher support has been cited as one reason that very promising technologies have failed to scale up and achieve sustained success" (Spector & Anderson, 2000, cited in Spector et al., 2016). Professional development and adequate resourcing of teachers enable them to respond to students with guidance as feedback creates the optimal learning experience (Spector, 2015).

Apart from the efficiency and practicality offered by intelligent tutoring systems, human teaching/tutoring involves confounding dynamics of relationship, credibility, mood, and perception. Learning results are not purely based on the experience of the process and a cognitive change in the subject. They are mitigated by past history with the instructor, the level of credibility that this individual has with the subject, the prevailing moods of both the teacher and the subject at the time of the assessment, and factors such as perceived judgment, skepticism, and support (Snow, 1986). In preparing an analysis of the effectiveness of human tutoring, VanLehn (2011) hypothesized that human tutors offer detailed diagnostic assessments, create individualized tasks, involve sophisticated strategies, allow learners to control the dialogue, have broad domain knowledge, inspire motivation, provide feedback, scaffold information, and follow an ICAP framework, which suggests the prioritization of interactive over constructive, constructive over active, and active over passive (p. 198-202). The *interaction granularity hypothesis* indicated that the key element differentiating human from computer tutoring is the level of granularity of attention (VanLehn, 2011). Critically important to the observation is the focus on step-based tutoring vs. answer-based tutoring. With the application of steps and the additional level of sub-steps, computer tutoring can approach the granularity of a human tutor and produce similar results (VanLehn, 2011). Results indicated that computer tutors in both the step-based and sub-step-based conditions yielded results equivalent to or better than human tutors, the researchers included a qualification of the results relating to the level of expertise of the human tutor (VanLehn, 2011).

Value of Quartile Analysis in providing targeted response

Student progress requires feedback on their performance. According to John Hattie:

> One of the ironies is that students who are above the average are less likely to ask for the 'what now?' feedback because they can usually work it out on their own. The kids who are below average really want the dialogue, want the information—and they're the least likely to get it. They get 'correct, incorrect, you could improve here'—checks and crosses that give them no information. (Sparks, 2018)

The needs of highly successful students and the needs of low-achieving students regarding feedback vary because struggling learners are not well-equipped to navigate the lessons, the assessments, and the decisions for themselves. Serving the middle or the median is a disservice to this subgroup. Determining a mean score holds little value for learners at either end of the achievement spectrum. While cluster analysis has shown some value in test-taking behavior implications, correlation coefficients reveal limited significance (Davis et al., 2008, p. 953). Quartile analysis, common in medical studies, appears less frequently in educational and behavioral analysis, where the mean, as a single descriptor of the whole, is the standard. It is in the more expansive and diverse analysis that variations, patterns, and behaviors come to the forefront.

With the *interaction granularity hypothesis* (VanLehn, 2011) and the *learning portrait model* (Peng et al, 2019), a method for unpacking the inner workings of the process of assessment and the decision-making characteristics of learners is possible, particularly when it utilizes computer-based systems which can be programmed to process and analyze the data, including quartile analysis and individual trends instantly.

Value of equipping educators for results analysis

According to the Atkinson theory of achievement motivation, learners possess a need for achievement and a need to avoid failure (Atkinson, 1978, cited by Snow, 1986). This intuitive explanation fuels the exploration of individual responses among students on an intelligent tutoring system or online assessment. The task cannot be viewed as a single performance, but is a series of individual performance events with independent potential and consequence. However, the study of learner performance in these environments has been characterized by a summary score of the final accuracy of the entire undertaking. Studies indicating mixed results can be the byproduct of ineffective targeting of the variables which should be measured. Each time a student attempts a response to a prompt, there is potential to achieve or to fail, accompanied by the need to achieve and the avoidance of failure (Atkinson, 1978, cited by Snow, 1986). The impact of the anxieties of the previous instance compound, amplifying the potential for a perpetuation of the direction, either toward correctness or toward error. The urgent need exists to explore these trends at the 'cellular' level, as recommended by the learning portrait model (Peng et al., 2019). Examination of students' responses for each cell, which includes a task and response, as well as the relationship of cell sequences, could be the key to unlocking the mystery of mixed results about feedback elements and understanding student performance.

Value of equipping students to overcome their negative test-taking patterns

Formative feedback is designed to "increase student knowledge, skills, and understanding of some content area or general skill...and there are multiple types of feedback that may be employed to this end" (Shute, 2008). It is directive feedback which targets items to revise, whereas facilitative feedback guides through suggestion (Shute, 2008). Results of the meta-analysis demonstrated that goal-directed feedback, which is specific, but not overly complex, and timely is most effective. If elaborative, delayed feedback can have stronger effects.

Research Questions

The following research questions will be explored in this analysis:

1. What patterns exist in accuracy and consistency between quartiles in test answer selection among eighth grade students?
    a. How do unsuccessful question attempts impact success on subsequent questions?
    b. How do successful question attempts impact success on subsequent questions?
2. Which quartile demonstrates the most consistent responses?

Method

Participants

This study was conducted at a midwest private, Catholic school in an urban setting. Eighth-grade students (*n*=61) accessed the application from September 2019 to May 2020 as a regular practice of their English coursework. The cohort of 27 males and 34 females ranged from 13- to 14-years old. Only assessment data during the year was used for calculating results. Participants were not contacted directly. Use of archived data constituted an exempt status for participant consent, although permission was given by both the classroom teacher and the school principal.

Assessments and Measures

All assessments were taken using a software application available at NoRedInk.com. The school licensed the software for improving grammar practice and assessment using diagnostics, practice lessons, and assessments. Students were asked to answer twenty questions with multiple answer choices available, of which one was correct. Assessment 1 (Commas) was given in October. Assessment 2 (Verbals) was given in December. Assessment 3, given in April, addressed active and passive voice. Most participants had experience with the software application prior to their first assessment, because of coursework during their seventh-grade year.

Answer coding.

Coding the responses for each NoRedInk quiz involved a series of strategies to map the behavior patterns. First, each answer was coded for accuracy to the question itself and the accuracy of the preceding and subsequent questions. For example, in a series of answers such as CNNCC, which reflects that the first question was answered correctly, the second and third answers were not correct, and that the fourth and fifth answers were correct, the code would begin with xCN (whereas "x" depicts no prior response), CNN, NNC, and NCC. The final code set ended in x to designate a null subsequent answer. In total, sixteen code variants were established -- XCC, XCN, XNC, XNN, CCC, CCN, CNN, CNC, NCC, NCN, NNC, NNN, CCX, CNX, NCX, and NNX. All codes were tallied for frequency by participant ID. Cross-references were tallied for each two-response pattern.

Quartiles.

Participants were grouped for each assessment based upon their final score. Groups were divided by quartile and labeled *Low*, *LowtoMid*, *MidtoHigh*, and *High*. Quartile determinations were based on an IBM SPSS analysis of the data for each assessment. Participants scoring in the identified range were labeled as indicated. In this case, quartile represents those participants whose scores are in the lowest quartile, not the lowest 25% (15.25 participants) because individuals in the bordering ranges with identical scores were placed in the same "quartile" for accuracy of numerical analysis.

Consistency/Inconsistency Groups.

The next step in coding analysis was to establish which patterns reflected consistent initial responses (CCC, CCN, CCX, NNN, NNC, NNX) and which indicated inconsistent initial responses (CNN, CNC, NCC, NCN, CNX, NCX). In the former instance, the accuracy or inaccuracy of the first response is duplicated on the second response. In the latter instance, the first response and second response are opposite of one another. The total number of codes demonstrating a consistent pattern was recorded for each participant, as was the total number of codes demonstrating an inconsistent pattern. The mean of the number of consistent patterns and the inconsistent patterns was calculated for each quartile and was reflected in the following tables with descriptive data (Tables 1-3).

Accuracy vs. Consistency.

While it may seem counterintuitive to analyze results by pattern of consistency instead of accuracy, this method promotes a less-biased representation of the data. Participants in the lowest quartile are predisposed to present incorrect answers more frequently than participants in the highest quartile. The opposite is true regarding correct answers. Therefore, identifying consistency instead of accuracy allows a comparison of each group in their area of strength. The correct and consistent (CC) condition and the incorrect consistent (NN) condition were aggregated in the Total Consistent calculation. The correct inconsistent (CN) and the incorrect inconsistent (NC) condition were aggregated in the Total Inconsistent calculation.

Results

Results are presented first as comparative means in chronological order of the test and then visually as stacked bar graphs of mean of initial response patterns by quartile for assessment. An analysis of variations over time will follow. Scores on Assessment 1 ranged from 30 to 100 percent, with 61 participants (*m*=74.59%; *sd*=15.229*)*. Scores on Assessment 2 ranged from 40 to 100 percent, with 60 participants (*m*=77.25%; *sd*=16.708). Scores on Assessment 3 ranged from 55 to 100 percent, with 59 participants (*m*=89.66%; *sc*=10.621). All assessments were conducted with the same participants, but the varying number of total participants reflected that some respondents opened the assessment, but did not answer any questions or answered only one question before ending the assessment. The data for these individuals was removed from the dataset when it seemed clear that analyzing their "incorrect" responses was not reflective of their knowledge, abilities, and choices, but of a different agenda in opening an assessment required by their teacher, but choosing not to engage in it.

Analysis of comparative mean data

When analyzing the results of the Assessment 1 data, a clear relationship was demonstrated between the level of consistency of answers and the quartile of the respondent (Table 1). The total of consistent responses increased gradually from the *Low* quartile to the *High* quartile. Tables 2 and 3 demonstrate rising performance scores with the same trend. Additionally, all tables demonstrated an inverse trend with the total of inconsistent responses.

Table 1. *Comparative mean data by quartile on Assessment 1(Commas Quiz).*

N=61		**Low**	**LowtoMid**	**MidtoHigh**	**High**
Total Consistent (CC & NN)	**Mean** **St. Dev.**	8.71 2.016	10.11 1.616	12.00 1.713	14.86 1.612
Total Inconsistent (CN & NC)	**Mean** **St. Dev.**	10.29 2.016	8.89 1.616	7.00 1.713	4.14 1.612
Correct Consistent (CC)	**Mean** **St. Dev.**	4.93 2.401	8.89 .928	11.13 1.204	14.77 1.602
Incorrect Consistent (NN)	**Mean** **St. Dev.**	3.79 2.424	1.22 .833	.88 .957	.09 .294

Correct Inconsistent (CN)	**Mean** **St. Dev.**	5.21 1.188	4.56 1.014	3.62 .957	2.14 .834
Incorrect Inconsistent (NC)	**Mean** **St. Dev.**	5.07 .997	4.33 .866	3.38 .885	2.00 .816

Table 2. *Comparative mean data by quartile on Assessment 2(Use of Verbals).*

N=61		**Low**	**LowtoMid**	**MidtoHigh**	**High**
Total Consistent (CC & NN)	**Mean** **St. Dev.**	11.21 3.068	12.14 1.460	14.00 1.683	16.89 1.560
Total Inconsistent (CN & NC)	**Mean** **St. Dev.**	7.79 3.068	6.86 1.460	5.00 1.683	2.11 1.560
Correct Consistent (CC)	**Mean** **St. Dev.**	5.93 2.814	10.43 1.089	13.23 1.301	16.89 1.560
Incorrect Consistent (NN)	**Mean** **St. Dev.**	5.29 2.091	1.71 .825	.77 .725	.00 .000
Correct Inconsistent (CN)	**Mean** **St. Dev.**	3.93 1.592	3.64 .842	2.54 .877	1.05 .780
Incorrect Inconsistent (NC)	**Mean** **St. Dev.**	3.86 1.512	3.21 .699	2.46 .877	1.05 .780

Table 3. *Comparative mean data by quartile on Assessment 3(Active and Passive Voice).*

N=61		**Low**	**LowtoMid**	**MidtoHigh**	**High**
Total Consistent (CC & NN)	**Mean** **St. Dev.**	14.45 2.162	15.47 1.179	17.00 .000	19.00 .000
Total Inconsistent (CN & NC)	**Mean** **St. Dev.**	4.55 2.162	3.53 1.179	2.00 .000	.00 .000
Correct Consistent (CC)	**Mean** **St. Dev.**	11.09 1.814	14.88 .857	17.00 .000	19.00 .000
Incorrect Consistent (NN)	**Mean** **St. Dev.**	3.36 2.248	.59 .712	.00 .000	.00 .000
Correct Inconsistent (CN)	**Mean** **St. Dev.**	2.27 1.104	1.82 .636	1.00 .000	.00 .000
Incorrect Inconsistent (NC)	**Mean** **St. Dev.**	2.27 1.104	1.71 .588	1.00 .000	.00 .000

Analysis of Increase Over Time in Mean of Initial Response Patterns

Figure 1 revealed a gradual increase in overall consistency from the *Low* to the *High* quartile in the height of the blue and red bands combined. A reduction of inconsistency from the *Low* to the *High* quartile was also indicated by the combined height of the green and orange bands. Consistent incorrect responses dropped between the *Low* and *LowtoMid* quartiles by 68%. The mean of the *MidtoHigh* quartile reflected a 77% reduction. The *High* quartile was 2% of the *Low* quartile mean. While the total of inconsistent responses diminished from the *Low* to the *High* quartile, they remained proportionate across both conditions (CN and NC). A similar pattern was demonstrated with Assessments 2 and 3.

Figure 1. *Stacked bar graph of mean of initial response patterns by quartile for Assessment1 1.*

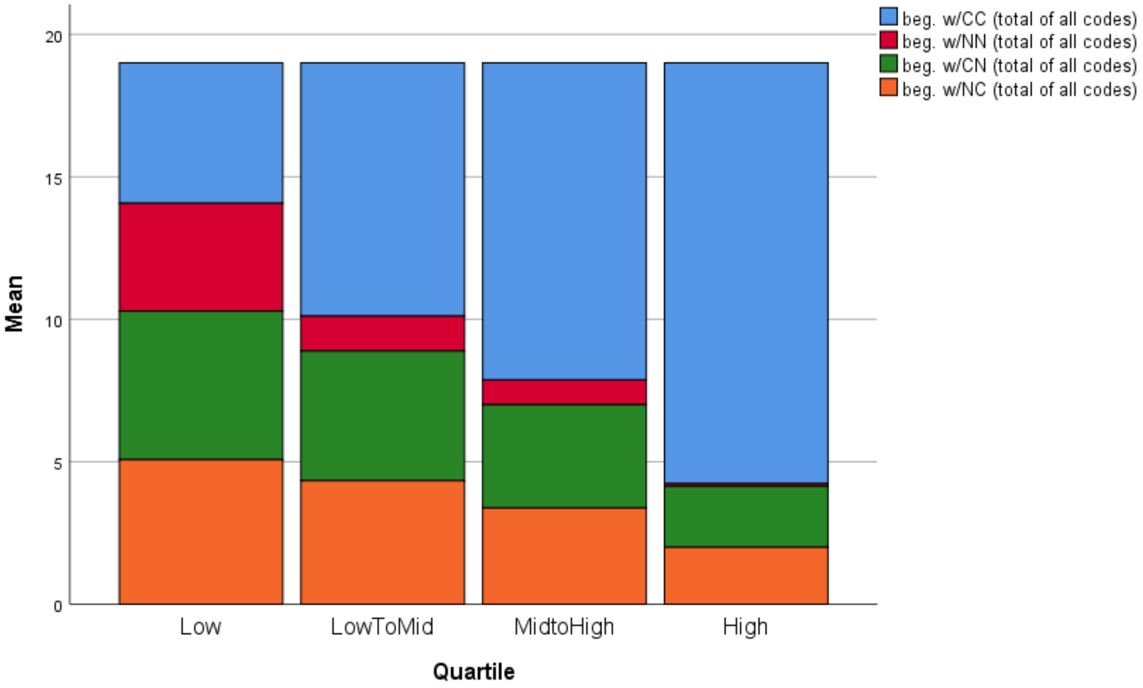


Figure 2. *Stacked bar graph of mean of initial response patterns by quartile for Assessment 2.*

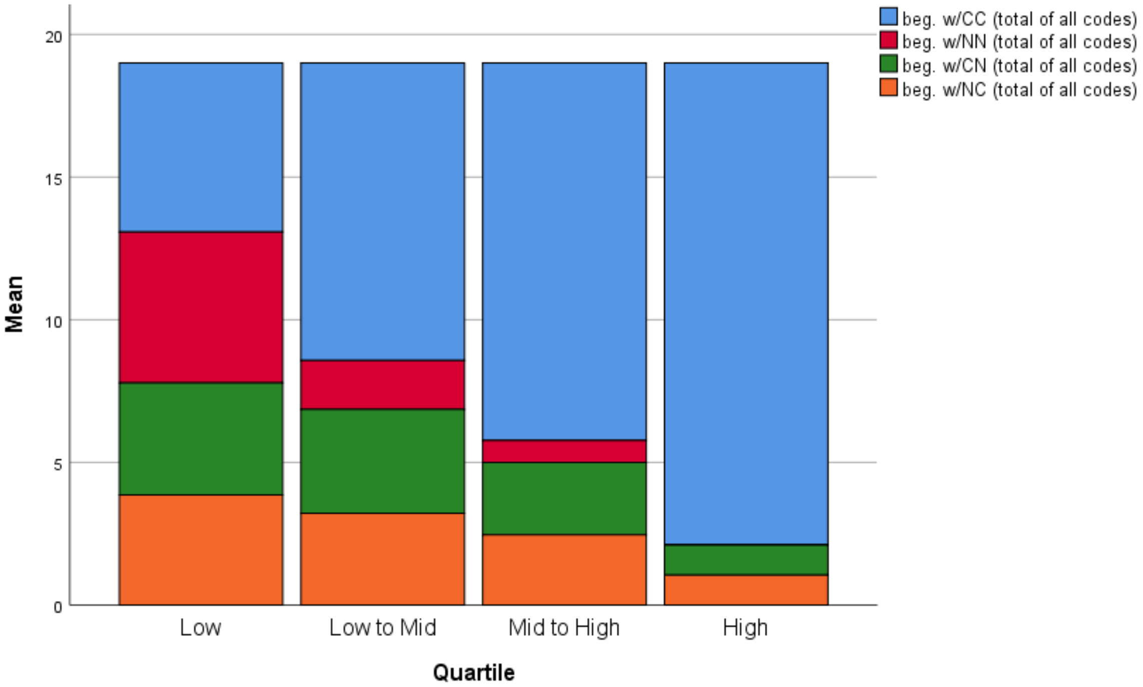
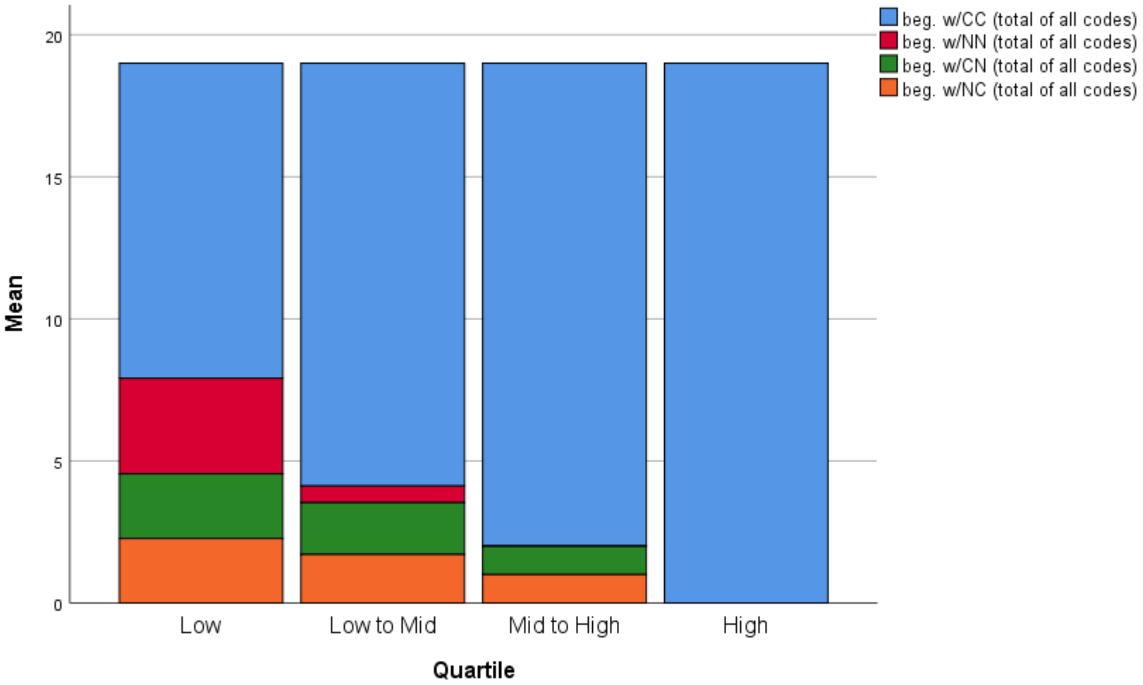


Figure 3. *Stacked bar graph of mean of initial response patterns by quartile for Assessment 3*



In order to investigate the patterns of consistency or inconsistency in greater detail, calculations were completed to assess the correlation between the Quartile variable and the Total Consistent/Total Inconsistent variables using the Pearson correlation method, Table 4 represents the correlations for Assessment 1, Table 5 for Assessment 2 and Table 6 represents Assessment 3.

Table 4. *Correlation of consistency level and accuracy with quartile for Assessment 1.*

N=61		Quartile	Total Consistent	Total Inconsistent	Correct Consistent	Correct Inconsistent	Incorrect Consistent
Total Consistent	**Pearson Sig (2-tailed)**	.814** .000					
Total Inconsistent	**Pearson Sig (2-tailed)**	-.814** .000	-1.000** .000				
Correct Consistent	**Pearson Sig (2-tailed)**	.917** .000	.904** .000	-.904** .000			
Correct Inconsistent	**Pearson Sig (2-tailed)**	-.810** .000	-.975** .000	.975** .000	-.898** .000		
Incorrect Consistent	**Pearson Sig (2-tailed)**	-.706** .000	-.388** .002	.388** .002	-.744** .000	.414** .001	
Incorrect Inconsistent	**Pearson Sig (2-tailed)**	-.779** .000	-.977** .000	.977** .000	-.867** .000	.905** .000	.345** .007

** Correlation is significant at the 0.01 level (2-tailed).

Table 5. *Correlation of consistency level and accuracy with quartile for assessment 2*

N=60		Quartile	Total Consistent	Total Inconsistent	Correct Consistent	Correct Inconsistent	Incorrect Consistent
Total Consistent	**Pearson Sig (2-tailed)**	.744** .000					
Total Inconsistent	**Pearson Sig (2-tailed)**	-.744** .000	-1.000** .000				
Correct Consistent	**Pearson Sig (2-tailed)**	.918** .000	.885** .000	-.885** .000			
Correct Inconsistent	**Pearson Sig (2-tailed)**	-.739** .000	-.990** .000	.990** .000	-.887** .000		

Incorrect Consistent	**Pearson Sig (2-tailed)**	-.818** .000	-.419** .001	.419** .001	-.794** .000	.436** .001	
Incorrect Inconsistent	**Pearson Sig (2-tailed)**	-.736** .000	-.991** .000	.991** .000	-.867** .000	.993** .000	.396** .002
** Correlation is significant at the 0.01 level (2-tailed).							

Table 6. *Correlation of consistency level and accuracy with quartile for Assessment 3*

N=59		**Quartile**	**Total Consistent**	**Total Inconsistent**	**Correct Consistent**	**Correct Inconsistent**	**Incorrect Consistent**
Total Consistent	**Pearson Sig (2-tailed)**	.818** .000					
Total Inconsistent	**Pearson Sig (2-tailed)**	-.818** .000	-1.000** .000				
Correct Consistent	**Pearson Sig (2-tailed)**	.935** .000	.832** .000	-.832** .000			
Correct Inconsistent	**Pearson Sig (2-tailed)**	-.812** .000	-.991** .000	.991** .000	-.844** .000		
Incorrect Consistent	**Pearson Sig (2-tailed)**	-.650** .000	-.251 .055	.251 .055	-.746** .000	.283* .030	
Incorrect Inconsistent	**Pearson Sig (2-tailed)**	-.809** .000	-.991** .000	.991** .000	-.806** .000	.965** .000	.215 .101
**** Correlation is significant at the 0.01 level (2-tailed); * Correlation is significant at the 0.05 level (2-tailed).**							

Discussion

While a performance total of 14/20 can be added to a gradebook, collected for a semester and reflect part of the final grade on a report card, an analysis resulting from Peng's Learning Portrait Model (2019) shares a far more illustrative portrayal of student answer patterns on a short online assessment. While a forthcoming study of students' persistence in correcting errors on this assessment addresses the need for feedback in greater detail, the assessment feedback offered to the teacher by this analysis provided valuable material for planning effectively, much like the outcomes of the Butler & Nisan study which indicated that grades yielded similar results to giving no feedback at all, but task-related feedback brought improvement (1986). The Learning Portrait Model (Peng et al., 2019) was feasible because of the structure of the NoRedInk app, through data-informed, digital formative assessment (Hwang, 2014; Spector et al., 2016). Armed with assessment results and the following analyses, educators can be better equipped to guide the needs of the learners in their charge.

Correlation of Quartile as a Measure

Participants in the lowest quartile demonstrated inverse response patterns from respondents in the highest quartile regarding the consistency of answer selections. Answers among respondents from the *Low* quartile were less consistent than participants in the *High* quartile. Pearson correlational analyses of between Quartile and Consistency for each assessment demonstrated a highly positive correlation at a highly significant level (p = .000). Assessment 1 yielded r = .814; Assessment 2 resulted in r = .744; and Assessment 3 revealed r = .818. The correlation between Quartiles and most measures of consistency and accuracy are strongly significant at each assessment timeframe, adding a high level of confidence to the validity of the results. The weakest correlations, though still evident, were demonstrated between incorrect consistent responses (NN) and both inconsistent conditions (CN and NC), suggesting that other factors, such as motivation, may contribute to patterns resulting in inaccurate responses.

Answering Research Questions: Impact of Success Level

The results of the mean analysis grouped by quartile (Figures 1, 2, and 3) paint the picture comprehensively. Accuracy increased in direct correlation to performance level. To be clear, students with low scores are predisposed to offer more incorrect answers than their peers with high scores, but the trend is not so simple. The decreasing trend by performance level, combined with assessment platform familiarity (detailed in 5.3) yielded an elimination of inaccuracy in Assessment 2 for the *High* quartile and for both the *MidtoHigh* and *High* quartiles in Assessment 3. Dramatic reductions in consistent but incorrect responses occurred between the *Low* and *LowtoMid* quartiles (68% on Assessments 1 & 2 and 82% on Assessment 3). However, the salient factors identified by the coded pattern analysis identified two other informative criteria: consistency/inconsistency and assessment platform familiarity.

Consistency

The analysis of consistency in the study addressed patterns of both correct and incorrect answers. Even the total of consistent correct responses combined with the total of consistent incorrect responses among learners in the *Low* quartile did not equal or exceed the same total for students in the *High* quartile. Students identified as unsuccessful based on the total test score (*Low* quartile), by definition, respond incorrectly more often than their successful peers in the *High* quartile.

Inconsistency

Participants in the *Low* quartile were more likely to respond with inconsistent patterns than those in the *High* quartile, at the highest level of statistical significance (p = .000). In short, students who score poorly are far more likely to demonstrate fluctuations in accuracy of response. Thus, it is the inconsistency itself, not the accuracy level, which reveals a trend. The benefit of this analysis provided clarity on the role of inconsistency as a trend among low-performing students, which can become a target for teacher strategy to guide them toward improvements.

Across all three assessments, data revealed that consistency was more important than accuracy. In Table 1, the correlation values were decisively high for students with correct and consistent responses, showing an inverse correlation to each other category. Interestingly, the strongest correlation among all four conditions was always between correct/inconsistent (CN) responses and incorrect/inconsistent (NC) responses. In assessment 1, *r*=.905; in assessment 2, *r*=.993; and in assessment 3, *r*=.965. This was illustrated by the green (CN) and orange (NC) bands on Figures 1, 2, and 3, which are proportionate and decline in magnitude at a slower rate regarding performance level. Therefore, while inaccuracy diminished and is eliminated in multiple conditions, inconsistency is present on all assessments at all performance levels except the Assessment 3 for the *High* quartile which was entirely correct. Not only is it present, but it is evident in the direct correlation at all performance levels, which underscores that the factor of inconsistency is more salient than the factor of accuracy.

Further, the lowest level of correlation across all assessments appeared between incorrect/inconsistent responses and incorrect/consistent responses. In Assessment 1, *r*=.345; in Assessment 2, *r*=.396; and in assessment 3, *r*=.215. It should be noted that the value for Assessment 3 did not rise to the 95% confidence level (*p*=.101), but also acknowledged that the accuracy rate for Assessment 3 was far higher than each of the other assessments. The low values for the correlation between types of incorrect responses reveals that it is not the strength of inaccuracy which drives the pattern, but the inconsistency of the responses which is most significant.

Platform Experience as an Assessment Barrier

Response pattern trends in inconsistency and inaccuracy diminished over the course of the assessment period from October 2019 to April 2020. The range of scores on Assessment 1 was 30 to100%, whereas Assessment 2 rose to 40 to100%, and Assessment 3 revealed 55 to100%, which shows the baseline increasing by 10% from Assessment 1 to Assessment 2 and increasing by 15% more between Assessments 2 and 3, with an overall baseline increase of 25% from the beginning to the end of the school year. Not only was the last assessment the best assessment for all groups, the comparative means between the first and second assessment demonstrated a trend of improvement in scores as well. The mean on Assessment 1 (*m*=74.59%) increased a slight 2% in two months by the time of Assessment 2, but increased a total of 15% (*m*=89.66%) by April. Even the standard deviation slimmed from 15.229 in October to 10.621 in April, underscoring the narrowing of the range and the focus of skills. Further, this evidence adds urgency to the importance of appropriate assessment and educational technology platform experience for learners. If familiarity with the platform is a barrier to successful responses for all students at early stages of experience, then providing adequate experience before students are subjected to a weighted assessment is a necessity.

Conclusion

While there are many factors which impact students when completing an online assessment of content knowledge, including anxieties, wakefulness, nutrition, visual processing, and dexterity, this study has revealed that a habit of consistency is also a strong influencer of successful outcomes.

Implications

Although few educators would be surprised by the statement that learners in the *Low* quartile are more likely to respond inconsistently on digital assessments or that students improve their performance over time as a function of experience in a digital environment, the evidence can be affirming. Armed with this data, teachers can provide guidance to all students in the introduction of the software application to the class. Proper introduction is likely to mitigate some of the issues which allow students to improve over time with hands-on experience through such explicit instructions as: 1) the goal of the task; 2) a vicarious, functional example; 3) a personal, live, and interactive experience with the application; and 4) an opportunity to see the results of the strategies employed so that the user can identify poor or strong practices, which can be discussed and used for later improvement. With this level of familiarity with the software application, students are more likely to be unencumbered by a lack of understanding or concerns that some of their actions may involve negative consequences.

As part of the effort to familiarize students with the software application, this study can assist teachers in helping to identify learners who may struggle with erratic decision-making during the time of testing. Learners who have not been explicitly taught to narrow a list of options, to reread the instructions for clarity, and to take additional care before making a final decision are more likely to perpetuate the model of inconsistent responses which are largely inaccurate. In the end, it is not about students making the right responses as much as it is about them answering authentically to the base of knowledge that they possess because it reveals a more accurate assessment of student knowledge than results which are flawed by behavioral anomalies and navigation-oriented mistakes.

Limitations

While this study provided a deep dive into the process of online assessment and the patterns of respondents in different quartiles, it was limited by its small population and application within subject matter. With only 61 participants, while it was an entire grade level for the school, the small group size limits generalizability to all eighth-grade students, all middle school students, and all grammar learners with the software used. Repeating the study over time with multiple cohorts of this grade level population in the school would lend itself toward rigor over time. Making a step toward a larger school system or toward an analysis of the archived data from the software application provider could offer some robust opportunities for assessing the validity of the trends in evidence from the current analysis. Additionally, similar data could be collected using other applications in other subject areas. Most applications which involve dashboards of student results to the teacher offer some levels of "drill down" reports which reveal an item-by-item analysis that could be retrieved and assessed for similar response patterns and trends.

Recommendations for Future Research

In addition to the explorations mentioned by means of overcoming limitations present in this study, the following data collection and analysis could be meaningful.

Gender effects. While the data set was not available for the current study, an analysis of the possible effects of gender as a contributing factor might yield intriguing results. Other studies have shown increased tendency for impulsivity in male participants (Fields, et al., 2009; Pham, 2016; Wise, et al., 2015), which could present similarly to the trend toward inconsistent answer patterns for the *Low* quartile of respondents in the present study. In fact, a combined follow-up of exploring additional grade levels along with gender effects could identify patterns related to age and gender at the same time.

Overall Performance Level. The use of a final accuracy score in the present assessment provides useful information regarding the level of the participant, but including data to show the respondent's final, comprehensive course grade as a grouping factor by quartiles might also lead to some effective comparisons. Situations in which the overall performance grade for the course differed markedly from the individual grammar assessment should shed light on the validity of scores. While the overall course score is likely to include a wider variety of components, such as composition skills, the integral nature of grammar in the subject area would seem to suggest a baseline of performance.

References

Bresciani, M. (2011). Identifying barriers in implementing outcomes-based assessment program review: A grounded theory analysis. Research & Practice in Assessment. 6(Summer), 5-16.

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. Journal of Educational Psychology, 78(3), 210-216. http://dx.doi.org.cmich.idm.oclc.org/10.1037/0022-0663.78.3.210

Claye, C. (1968). Barriers to Effective Teaching. *The Journal of Negro Education, 37*(2), 146-152.

Davis, H. & DiStefano, C. & Schutz, P. (2008). Identifying patterns of appraising tests in first-year college students: Implications for anxiety and emotion regulation during test taking. *Journal of Educational Psychology. 100*(4), 942-960. DOI:10.1037/a0013096.

DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling Student Cognition in Digital and Nondigital Assessment Environments. *Educational Assessment, 22*(4), 275–297. https://doi-org.cmich.idm.oclc.org/10.1080/10627197.2017.1382343

Fields, S., Collins, C., Leraas, K., & Reynolds, B. (2009). Dimensions of impulsive behavior in adolescent smokers and nonsmokers. *Experimental and clinical psychopharmacology, 17*(5), 302.

Gallo, A., Sheehy, D., Patton, K., & Griffin, L. (2006). Assessment Benefits and Barriers. *Journal of Physical Education, Recreation & Dance, 77*(8), 46-50.   https://doi-org.cmich.idm.oclc.org/10.1080/07303084.2006.10597926

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. Computers in Human Behavior 61, 36-46. https://doi.org/10.1016/j.chb.2016.02.095

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. Retrieved from http://cmich.idm.oclc.org/login?url=https://search-proquest-com.cmich.idm.oclc.org/docview/214113991?accountid=10181

Hwang, G. J. (2014). Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective. *Smart Learning Environments, 1*(1), 4.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254. https://pdfs.semanticscholar.org/97cc/e81ca813ed757e1e76c0023865c7dbdc7308.pdf

Koedinger, K., Brunskill, E., Baker, R., McLaughlin, E., & Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine, 34*(3), 27-41.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of Feedback and Verbal Learning. *Review of Educational Research, 58*(1), 79–97. https://doi.org/10.3102/00346543058001079

Paassen, B., Mokbel, B., & Hammer, B. (2016). Adaptive structure metrics for automated feedback provision in intelligent tutoring systems. *Neurocomputing, 192*(C), 3-13.

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of Solution Behavior across Items in Low-Stakes Assessments. *Educational Assessment, 24*(3), 189–212. https://doi-org.cmich.idm.oclc.org/10.1080/10627197.2019.1615373

Peng, H., Ma, S. & Spector, J.M. (2019). Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, *6*(9) doi:10.1186/s40561-019-0089

Pham, A. V. (2016). Differentiating behavioral ratings of inattention, impulsivity, and hyperactivity in children: effects on reading achievement. *Journal of attention disorders, 20*(8), 674-683.

Shute, V. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*(1), 153-189.

Slof, B., Erkens, G., Kirschner, P., and Helms-Lorenz, M. (2013). The effects of inspecting and constructing part-task-specific visualizations on team and individual learning. Computers & Education 60, 221-233. https://www-sciencedirect-com.cmich.idm.oclc.org/science/article/pii/S0360131512001613

Snow, R. (1986). Individual Differences and the Design of Educational Programs. *American Psychologist, 41*(10), 1029-1039.

Sparks, S.D. (2018). Getting feedback right: A Q&A with John Hattie. *Education Week 37*(36), 8-9. https://www.edweek.org/ew/articles/2018/06/20/getting-feedback-right-a-qa-with-john.html

Spector, J. M. (2015). *Foundations of educational technology: Integrative approaches and interdisciplinary perspectives* (2nd ed.). New York: Routledge.

Spector, J. M. (2016). The potential of smart technologies for learning and instruction. *International Journal of Smart Technology and Learning, 1*(1), 21-32.

Spector, J. M., & Anderson, T. M. (Eds.) (2000). *Integrated and holistic perspectives on learning, instruction and technology: Understanding complexity.* Dordecht, Netherlands: Khwer Academic Press.

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, J. L., Mukama, E., Warusavitarana, A., ... & Bridges, S. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of Educational Technology & Society, 19*(3), p. 58-71.

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist, 46*(4), 197-221. DOI:10.1080/00461520.2011.611369

Wise, R. J., Phung, A. L., Labuschagne, I., & Stout, J. C. (2015). Differential effects of social stress on laboratory-based decision-making are related to both impulsive personality traits and gender. *Cognition and Emotion, 29*(8), 1475-1485.

Xie, L. (2018). Research of Computer-aided Instruction based on Virtual Reality Technology. *2018 3rd International Conference on Education and Education Research.*

Yang, L., Sin, K. F. K., Li, X., Guo, J., & Lui, M. (2014). Understanding the power of feedback in education: A Validation study of the Feedback Orientation Scale (FOS) in classrooms. International Journal, 16, 1. https://www.researchgate.net/profile/Lan_Yang17/publication/290404660_Understanding_the_Power_of_Feedback_in_Education_A_Validation_Study_of_the_Feedback_Orientation_Scale_FOS_in_Classrooms/links/56970a4b08ae34f3cf1df590.pdf