Georgia Educational Researcher

Volume 12 | Issue 2 Article 1

12-31-2015

Analyzing Mathematics and AP Statistics Assessment Items in Terms of the Levels of Thinking Skills They Assess

Shadreck S. Chitsonga

Analyzing Mathematics and AP Statistics Assessment Items in Terms of the Levels of Thinking Skills They Assess

Abstract
This study is a part of a large study that investigated the Assessment Practices of Mathematics Teachers who also teach AP Statistics. In this paper I investigate the thinking skills that assessment items used mathematics and AP Statistics classroom assess. Two teachers who were teaching mathematics and AP Statistics participated in this study. Each teacher was observed 12 times in class. Artifacts were also collected from each teacher, mainly oral questions and written questions. A mathematics taxonomy framework was used to analyze the characteristics of the assessment items. The results of the study indicated that assessment questions (oral and written) in AP Statistics and mathematics mostly assessed recall of factual information, comprehension of factual information and routine use of procedures.

Keywords
Mathematics, AP Statistics, Thinking Skills, Assessment Items, Mathematics Taxonomy Framework

Analyzing Mathematics and AP Statistics Assessment Items in Terms of the Levels of Thinking Skills They Assess

Shadreck S. Chitsonga
Fort Valley State University
Fort Valley, GA

Abstract: This study is a part of a large study that investigated the Assessment Practices of Mathematics Teachers who also teach AP Statistics. In this paper I investigate the thinking skills that assessment items used mathematics and AP Statistics classroom assess. Two teachers who were teaching mathematics and AP Statistics participated in this study. Each teacher was observed 12 times in class. Artifacts were also collected from each teacher, mainly oral questions and written questions. A mathematics taxonomy framework was used to analyze the characteristics of the assessment items. The results of the study indicated that assessment questions (oral and written) in AP Statistics and mathematics mostly assessed recall of factual information, comprehension of factual information and routine use of procedures.

Keywords: Mathematics, AP Statistics, Thinking Skills, Assessment Items, Mathematics Taxonomy Framework

Analyzing Mathematics and AP Statistics Assessment Items in Terms of the Levels of Thinking Skills They Assess

Introduction

Tasks are important in the learning of mathematics—they send messages to students about what mathematics is and what is involved in doing mathematics (NCTM, 1991). Some researchers have commented on the creation of assessment tasks. For example, Senk, Beckmann, and Thompson (1997) have proposed that

> future research investigate effective models for preservice and in-service education to help teachers address issues in assessment and that future curriculum development efforts include the creation of assessment tasks. In particular, teachers need examples of worthwhile assessment tasks geared to specific courses they are teaching, rather than examples that are meant to assess general levels of mathematical performance. (p. 210)

It is the responsibility of teachers to choose quality mathematics tasks that students can engage in (NCTM, 1991). The NCTM Task Standard defines good tasks as

> ones that do not separate mathematical thinking from mathematical concepts or skills that capture students' curiosity and that invite them to speculate and to pursue their hunches. Many of such tasks can be approached in more than one interesting and legitimate way; some have more than one reasonable solution. These tasks, consequently, facilitate significant classroom discourse, for they require that students reason about different outcomes, weigh the pros and cons of alternatives, and pursue particular paths. (p. 25)

DOI: 10.20429/ger.2015.120201

One way of determining the quality of assessment tasks is to determine the levels of thinking skills that these tasks assess.

Background

The inclusion of data analysis and probability in the National Council of Teachers of Mathematics (NCTM) 1989 and 2000 standards documents has been a welcome development insofar as the recognition of the importance of statistics in the PreK–12 curriculum is concerned. Apart from the NCTM, a number of organizations have undertaken projects to improve the teaching of statistics in Grades PreK–12. Among those is the project on Guidelines for Assessment and Instruction in Statistics Education (GAISE; Franklin et al., 2007), which has been recognized as a welcome development by the American Statistical Association (ASA, 2005).

Endorsing GAISE, the ASA said:

> The ASA Board of Directors appreciates and supports the efforts of the Pre K–12 GAISE writing group and endorses its recommendations for Pre K–12 statistical education in the document, *A Curriculum Framework for Pre K–12 Statistics Education* as an aid to enhancing statistics education at the K–12 levels. (p. 1)

The GAISE Report (Franklin et al., 2007) gives a clear picture of the status of statistics education in the secondary school in terms of the teachers' knowledge of statistics:

> Statistics . . . is a relatively new subject for many teachers, who have not had an opportunity to develop sound knowledge of the principles and concepts of data analysis that they are now called upon to teach. These teachers do not clearly understand the difference between mathematics and statistics. (p. 5)

It is important to note that statistics does not originate from within mathematics. Statistics is concerned with gathering, organizing, and analyzing data, and with inferences from data to the underlying reality (Moore, 1988, p. 4). The distinction between mathematics and statistics may seem trivial. But the fact that some teachers cannot make the distinction between the two is a huge problem because those teachers teach statistics as a mathematical topic, putting a lot of emphasis on computations, formulas, and procedures (Gal & Garfield, 1997). This has an implication on the choice of assessment tasks that teachers use.

Rationale

Though there is plenty of literature on assessment practices of secondary school teachers of mathematics, there is no study that has focused on the secondary teachers of mathematics who also teach statistics. The fact that many teachers of mathematics are being called upon to teach statistics means that studies related to the teaching of statistics deserve serious consideration. More and more statistics topics are finding their way into the K–12 mathematics curriculum as calls to integrate statistics with other subject areas become more numerous. The GAISE framework suggests the following: "The traditional mathematics strands of algebra, functions, geometry, and measurement all can be developed with the use of data. Making sense of data should be an integrated part of the mathematics curriculum, starting in pre-kindergarten" (Franklin et al., 2007, p. 35). With all these developments, it is also imperative to examine the assessment practices of these teachers who are teaching both mathematics and statistics. Specifically, the present study looks at the thinking skills that assessment tasks (items) in mathematics and AP statistics assess.

Research Questions

This study aims to address the following research questions:

DOI: 10.20429/ger.2015.120201

1. Which levels of thinking skills do assessment tasks used in mathematics classrooms assess?
2. Which levels of thinking skills do assessment tasks used in AP Statistics classrooms assess?
3. Are there any differences in thinking skills assessment between oral assessment items and written assessment items used in mathematics and AP Statistics classes?
4. Are there any differences in thinking skills assessment based on subject area?

Literature Review

Stiggins, Griswold, and Wikelund (1989) examined the classroom assessment procedures of 36 teachers in Grades 2 to 12 to determine the extent to which the teachers in mathematics, science, social studies, and language arts measure higher order thinking skills in their respective subject areas. Stiggins et al. used a framework of thinking skills developed by Quellmalz (1985). The framework includes five types of thinking skills: recall, analysis, comparison, inference, and evaluation. Recall is the lowest-level, and evaluation is the highest level. They reported that in all the subjects but mathematics, the written assessment questions had a heavy reliance on recall items. In mathematics, the study found that 72 % of the items assessed inference, and only 19% of the items assessed recall. This result implies that only mathematics assessed high-level thinking skills. However, this result contradicts with the findings of Duncan and Noonan (2010). In their study with 513 high school teachers, they found that in all subject areas (math, English, social studies , practical arts, and religious studies), no significant differences were found in terms of the cognitive levels of assessment. The emphasis of high order thinking (measuring

understanding, reasoning, and application) was almost the same in all the subject areas. In other words, subject area did not have any influence on the thinking skills assessed.

Stiggins et al. (1989) also observed that items assessing other thinking skills such as comparison and evaluation were rarely used. They suggested a reason that evaluation is largely ignored:

> Evaluation may be viewed as difficult to address because there may be no right answer. Teachers may not feel secure enough in the subjects they teach to be able to guide and in fact evaluate answers to evaluation questions in terms of the strength of the defense provided for an opinion. But in general, both comparison and evaluation are very important thinking skills that need to be developed and assessed. (p. 243)

One other interesting finding of the Stiggins et al. 1989 study was that there was a remarkable difference between oral questions and written questions in mathematics: Written assessments were predominantly inference, and oral questions were distributed evenly across recall, analysis, and inference. A similar observation was made by Delice, Aydin, and Cevik (2013). They studied mathematics teachers' use of questions. Specifically, they looked at the factors influencing preparation of questions. They reported that during teacher-student interactions, [oral] questions were generally asked to help students realize 'connections between knowledge' and to increase motivation. However, teachers gave importance to the development of higher mental processes in their [written] homework questions.

Suah and Ong (2012) investigated the assessment practices of 406 teachers. They reported that when developing tests, the teachers did not prepare a table of specifications to help them determine the cognitive levels of the items. This implies that the teachers did not think

DOI: 10.20429/ger.2015.120201

about the cognitive levels of the assessment items they developed. Suah and Lan also reported that teachers with less experience were unable to construct their own test questions. They attributed this to lack of necessary skills to develop quality test items.

Senk et al. (1997) studied the assessment and grading practices of high school teachers of mathematics in Algebra, Geometry, Advanced Algebra, Trigonometry, and Functions and Precalculus. They reported that in all the courses, the tests had a high percentage of low-level items (ranging from 53 to 90%). However, this finding is contrary to what Stiggins et al. (1989) found (as indicated above). Senk et al. also found that there was not enough evidence to suggest that the test items required the students to use reasoning (justification, explanation, or proof). The mean usage was 5%, the highest percentage of the usage was by the teachers of Geometry courses, and there was no evidence that teachers of Algebra I or II courses used the items. Senk et al. (1997) also noted that the teachers seldom or never used any open-ended items on tests (percentages of open-ended items ranged from 0 to 10%). Sanchez (2002) reported a different result about the use of open ended assessments. However, she outlined the reasons why the teachers in her study used open-ended assessments:

> The teachers in this study each used open-ended items that assessed higher-level thinking skills on every test. That was likely due to the teachers' participation in the assessment projects and the support they received from their school system for using open-ended items. Too, the system actually required that they use open ended items on tests, so the teachers' orientation to authority also affected their use of open-ended items on tests. (p. 120)

Method

This study consisted of two case studies. I considered the case study design appropriate for this study because it gave me the opportunity to look closely at, understand, and report on the assessment practices of mathematics teachers who also teach AP Statistics. Furthermore, because there have not been any studies on assessment that focus on those mathematics teachers who also teach AP Statistics, a case study was ideal to gain a deeper understanding of the assessment practices of the teachers involved.

Participants

The participants in this study were two high school teachers named John and Mary (pseudonyms). At the time of the study, the participants were teaching both mathematics and AP Statistics. The Advanced Placement (AP) Program offers a course description and exam in statistics to secondary school students who wish to complete studies equivalent to a one-semester, introductory, non-calculus-based, college course in statistics (College Board, 2007, p. 3). Both were members of a local professional learning community for AP Statistics. The AP Statistics professional learning community was a gathering of AP Statistics teachers in a Southeastern state of the United States of America. The main objective of the gathering was to help AP Statistics share ideas on the teaching of statistics. The participants were identified based on purposeful sampling. According to Patton (2002), the power of purposeful sampling lies in the selection of information rich cases. He says, "Studying information rich samples yields insights and in-depth understanding rather than empirical generalizations" (p. 230).

Procedure

I consulted with the coordinators of the professional learning community for AP Statistics teachers to assist in the identification of participants who had experience in teaching both

DOI: 10.20429/ger.2015.120201

mathematics and AP Statistics. The teachers identified were members of this learning community for AP Statistics teachers. Their participation in the learning community was used as a criterion of selection for them to take part in this study. Once the participants were identified, I went to their schools to meet them and brief them about the study. They were both gracious and agreed to participate in the study. I also asked them if they could allow me to sit in their classes before beginning the study so that I could familiarize myself with the school environment. They agreed to do that, and I observed each teacher a number of times in his or her mathematics and AP Statistics classes. During that time I did not collect any data.

Data Collection Procedures

Data for the study were collected through multiple methods. In this study, three methods were used to collect data: semi-structured personal interviews, observations, and collection of artifacts (documents like quiz and test papers).

The two participants were each interviewed two times. In the first interview, I asked the participants to give their background information, such as their educational background, with a special focus on their preparation to become teachers. The second interview addressed issues related to assessment in both mathematics and AP Statistics. The questions asked in the interviews were the same for both teachers. The interviews were recorded using a digital audio recorder. Additional informal interviews were also conducted at the end of each class observation whenever necessary—there were nine in total, four for John and five for Mary. These interviews were mainly to seek clarifications of observations made in class. These interviews were generally very brief, as immediately after the observation another class was coming into the room, so there was not enough time to discuss many things. Sometimes I obtained clarification through an exchange of emails with the participants.

I also made classroom observations. I observed each teacher 12 times (6 times in the mathematics class and 6 times in the AP Statistics class). Each observation lasted the entire teaching period, which was 90 minutes for both mathematics and AP Statistics. During the observations, I made field notes to record what I observed. I observed the type of questions the teachers asked in class.

Description of the Schools

The two schools used in the study were both in a Southeastern state county system in the United States of America. They were the only two public high schools in the county. The school year for both schools consisted of two 18-week terms, and the schools operated on a 4×4 modular or block schedule. All the classes were 90 minutes in duration.

John's School

John's school had a population of slightly over 1000 students. About 90% of these students were White. The remaining 10% were Blacks, Asian, Hispanic, or American Indian/Alaskan.

John's classroom was large enough to accommodate 30 desks and chairs. John's class had extra material such as a set of textbooks that students could use if they had not brought their textbook. Also available was a set of calculators and graphing utilities that students could borrow to use in class, though most had their own calculator.

At the time of the study, John was teaching two classes of AP Statistics, one class of Mathematics Support I, and one class of Precalculus. I observed John in one AP Statistics class and his Precalculus class. John's AP Statistics class had 15 students, all of whom were White; there were 7 boys and 8 girls. John's Precalculus class had 19 students, 11 of whom were boys. Three of the students were Black, and the rest were White.

DOI: 10.20429/ger.2015.120201

Mary's School

The school's student population was slightly less than 1000 students. Ninety percent of the students were White; the remaining 10% were Black, Asian, or Hispanic. The gender ratio in the student population was almost 1:1.

Mary's classroom had 28 desks and chairs. The classroom had overhead projector that Mary used most of the time while teaching. She also had some extra books that were used by the students if they had forgotten their textbook. Other materials in the room included a set of TI-83 graphing calculators and ordinary calculators that could be lent to students who did not have their own or had forgotten to bring theirs to class.

At the time of the study, Mary was teaching two classes of Honors Algebra II, one class of Mathematics Support I, and one class of AP Statistics. I observed Mary teaching one Honors Algebra II class and the AP Statistics class. Mary's AP Statistics class had 17 students, all of whom were White with an exception of one Black student. Eight of the AP Statistics students were girls; the Honors Algebra II classes had 18 students, 8 of whom were girls. All the students in Honors Algebra II were White.

Mary taught all her classes in the same room, with an exception of lessons that required the use of computers. During the time of the study, Mary took her AP Statistics class to the computer laboratory once.

Data Analysis

The present study examined the type of tasks teachers use in their mathematics and AP Statistics classrooms. Specifically, I looked at the thinking skills that those tasks assess. These tasks comprised the oral questions that the teachers used in class and the written assessment tasks, which were mainly, quiz and test items. These tasks were analyzed by using the

mathematics taxonomy (Table 1) developed by Smith, Wood, Compland, and Stephenson (1996), which is a modification of the 1956 Bloom taxonomy.

Table 1
Taxonomy of Student Understanding in Mathematics

Group A	Group B	Group C
Factual knowledge	Information transfer	Justifying and interpreting
Comprehension	Application to new situations	Implications, conjectures, and comparisons
Routine use of procedures		Evaluation

Note. From Smith et al. (1996, p. 67).

The Smith et al. taxonomy was developed to make it compatible with the purpose of assessing students' understanding in mathematics. I chose to use this taxonomy for a number of reasons:

- It was specifically designed for analyzing mathematical tasks.
- The classification of the tasks in terms of thinking skills is in line with NCTM's (1989) *Curriculum and Evaluation Standards for School Mathematics*. For example, Standard 3 (mathematical reasoning) stipulates that students should "make and test conjectures, formulate counterexamples, [and] construct proofs for mathematical assertions, including indirect proofs and proofs by mathematical induction" (p. 143).
- Though it was constructed specifically for mathematics, the taxonomy is also applicable to statistics tasks. Of all the frameworks I looked at, this was the best for analyzing both mathematics and statistics tasks.

The taxonomy has three main groups—namely, A, B, and C—encompassing eight categories. Group A is the first level of thinking, and its tasks require students to recall "previously learned information in the form it was given" (Smith et al., 1996, p. 68), or work

DOI: 10.20429/ger.2015.120201

with problems or tasks they have had plenty of practice with in class. Group B tasks ask students to use their knowledge beyond simple recall or working with routine problems. They require students to use their knowledge in unfamiliar or new situations. In Group C tasks, students must demonstrate abilities and skills like justifying, interpreting, comparing, and evaluating. Examples of the tasks for each group appear in Appendix.

Results

Analysis of Oral and Written Assessment Items Used in John's Mathematics class

Table 2 presents results of my coding of the oral and written assessment items used by John in his Precalculus class. The results show that most of the oral and written assessment items were in Group A (first level of thinking). It is important to note, however, that the written assessment items did not assess any recall of factual knowledge. For example, students were never asked to state a theorem, whereas in oral assessment, students were at times asked to do that. Other thinking skills—that is, those in Groups B and C—were seldom assessed either in oral or written assessment items.

Table 2
Percent of John's Oral and Written Assessment Items Assessing Thinking Skills in Mathematics

Group	Description	Oral questions	Test and quiz items
Group A	Factual knowledge	27	0
	Comprehension	43	23
	Routine procedures	27	71
Group B	Information transfer	0	2
	Application in new situations	0	0
Group C	Justifying and interpreting	0	4
	Implications, conjectures and comparisons	3	0
	Evaluation	0	0

Further analysis of the written assessment items showed that none of the items were open ended. The majority of the questions, over 90%, were open-middled. According to Bush and Greer (1999), open-ended items are those that have multiple solutions and whose solutions can be found in several ways. Open-middled items are those that have one correct answer, but there are multiple ways of getting to that answer. There was no clear rationale for the use of these items. However, I noted that John used multiple-choice items in quizzes and not in tests. The small percentage of multiple-choice items could be attributed to the fact that John did not like such items, as he indicated that he liked tests/quizzes that required students to investigate, write and "you know, it is more than just a one-question thing." But it is also possible that John just used the format in the textbook publisher's test book. The same could be said of the absence of open-ended items in the quizzes and tests.

Analysis of Oral and Written Assessment Items Used in John's AP Statistics Class

Table 3 presents results of the coding of oral and written assessment items used by John in his AP Statistics class.

Table 3
Percent of John's Oral and Written Assessment Items Assessing Thinking Skills in AP Statistics

Group	Description	Oral questions	Test and quiz items
Group A	Factual knowledge	27	25
	Comprehension	14	24
	Routine procedures	43	29
Group B	Information transfer	5	3
	Application in new situations	0	0
Group C	Justifying and interpreting	11	15
	Implications, conjectures and comparisons	0	7
	Evaluation	0	0

DOI: 10.20429/ger.2015.120201

The results show that most of the assessment items, both oral and written were in Group A. In the written assessments, the percentage of the items requiring factual knowledge, comprehension of factual information, and routine use of procedures were not that much different, whereas for the oral assessment items, the majority of the items in Group A required comprehension of factual knowledge. Items requiring thinking skills in Group B were hardly used. The percentage for written assessment items in Group C was twice that of the oral assessment items. It was evident, however, that posers of both oral and assessment items ignored the application in new situation and evaluation thinking skills.

Further analysis of the written assessment items showed that less than 2% of the items were open ended. In general, there were no multiple-choice or true-false items in the quizzes. In the tests, however, almost 50 % of the items were multiple-choice. John used such items so as to conform to the structure of the AP Statistics Examination.

Analysis of Oral and Written Assessment Items Used in Mary's Mathematics Class

Table 4 presents results of the coding of oral and written assessment items used by Mary in her Honors Algebra II class.

Table 4
Percent of Mary's Oral and Written Assessment Items Assessing Thinking Skills in Mathematics

Group	Description	Oral questions	Test and quiz items
Group A	Factual knowledge	27	0
	Comprehension	14	6
	Routine procedures	43	94
Group B	Information transfer	5	0
	Application in new situations	0	0
Group C	Justifying and interpreting	11	0
	Implications, conjectures, and comparisons	0	0
	Evaluation	0	0

Most of the assessment items for both oral and written assessments were in Group A. Almost all the written assessment items required students to use routine procedures. Eleven percent of the oral assessment items were in Group C, but there were no written assessment items assessing thinking skills in either Group B or Group C. Clearly, the results show that written assessment items were dominated by items that required the use of routine procedures. This dominance was to be expected because most of the items used in the quizzes or tests were similar to those used in homework assignments or discussed in class.

A further look at the written assessment items revealed that none of the items was open ended even though Mary had indicated that she liked open-ended items. It would appear that Mary might have viewed any open-middled questions as open ended. All the items in the tests were open-middled. In the quizzes, there was a mixture of multiple-choice, true-false, and open-middled items. Mary did not seem to have criteria to determine the number of items of a certain format in the quiz; for example, in one quiz 85 % of the questions were multiple choice, whereas in another quiz only 23% of the questions were multiple choice.

Analysis of Oral and Written Assessment Items Used in Mary's AP Statistics

Table 5 presents results of the coding of oral and written assessment items used by Mary in her AP Statistics class.

DOI: 10.20429/ger.2015.120201

Table 5
Percent of Mary's Oral and Written Assessment Items Assessing Thinking Skills in AP Statistics

Group	Description	Oral questions	Test and quiz items
Group A	Factual knowledge	17	2
	Comprehension	26	11
	Routine procedures	43	78
Group B	Information transfer	0	4
	Application in new situations	0	0
Group C	Justifying and interpreting	9	4
	Implications, conjectures and comparisons	5	0
	Evaluation	0	0

An analysis of the assessment items showed that the majority of them required the use of routine procedures. The percentage for written assessment items, however, was higher than that of the oral assessment items. Items in Groups B and C were almost nonexistent in the written assessments, with only 4% in each category. For oral assessments, there were no items in Group B, whereas the percentage of items in Group C was 14%. Though small, this percentage was more than three times that of the written assessment items.

I analyzed the written assessment items further to determine whether they were open ended or not. The results indicated most the items (over 95%) were not open ended. The only items that could be described as open ended involved the designing of simulations. Half of the items in the tests were multiple choice. However, there were no multiple-choice items in the quizzes.

Cross Case Analysis Oral Questions

Comparisons of the oral questions the teachers asked revealed that both teachers relied on assessment items that required recall of factual information, comprehension, and use of routine procedures in mathematics and AP Statistics. For both teachers, over 80% of the oral items in

mathematics and AP Statistics assessed Group A thinking skills (items that require recall of factual information, comprehension, and the use of routine procedures; Smith et al., 1996). In terms of subject matter, there appeared to be no difference between the two teachers in the use of items assessing Group A thinking skills. There was a heavy emphasis on the use of these items in both subjects. Items assessing Group B thinking skills (items that require information transfer and application in new situations) were largely absent in mathematics and AP Statistics for both teachers; the percentages of such items ranged between 0 and 5. And in the case of the items in Group C (items require justifying, interpreting, comparing and evaluating), subject matter did not make any difference. For Mary, percentages were small and almost identical in both subjects. In contrast, John did not use items requiring the skills in Group C in mathematics, but he did in AP Statistics, which indicated that subject matter determined whether he used questions measuring thinking skills in Group C.

Cross Case Analysis of Written Assessment Items

For both teachers, the items used in the tests and quizzes in mathematics and AP Statistics were dominated by Group A thinking skills. The percentage of the items in this category ranged between 78 and 100%. There appeared to be no difference in terms of subject matter in the use of questions assessing Group A thinking skills. Items assessing Group B thinking skills were used minimally in both mathematics and AP Statistics by both teachers, with the percentages 4% or below. In John's case, the analysis indicated that he was more likely to use items requiring Group C thinking skills in AP Statistics than in mathematics (22% of the written AP Statistics items were in Group C compared with only 4% in mathematics). The subject matter determined the way he used items requiring Group C thinking skills.

DOI: 10.20429/ger.2015.120201

Discussion

The purpose of this study was to investigate the thinking skills assessed by assessment tasks used in mathematics and AP Statistics. I used the mathematics taxonomy framework by Smith et al. (1996) to analyze the tasks the teachers used for assessment. The analysis focused on questions. In the study, I addressed four major research questions:

1. Which levels of thinking skills do assessment tasks used in mathematics classrooms assess?
2. Which levels of thinking skills do assessment tasks used in AP Statistics classrooms assess?
3. Are there any differences in thinking skills assessment between oral assessment items and written assessment items used in mathematics and AP Statistics classes?
4. Are there any differences in thinking skills assessment based on subject area?

Previous studies (Senk et al., 1997; Stiggins et al., 1989) have indicated that teachers ask mostly low-level questions. The present study yielded the same conclusion. The teachers in this study used oral and written assessment items that required a low level of thinking; namely, recall of factual information, comprehension, and use of routine procedures. This was the same in both mathematics and AP statistics. As in Stiggins et al.'s study, the teachers in the present study ignored other thinking skills such as applications in new situations, justifying, and interpreting. Stiggins et al suggested that it might be difficult for teachers to guide students and evaluate answers in evaluation questions. I did not establish in this study that teachers had the same problems with evaluation questions as these suggested by Stiggins et al. (1989). There were situations, however, when the teachers showed the ability to evaluate students' answers that were

different from what they expected. The absence of items assessing higher levels of thinking could indicate that the teachers never thought about classifying the assessment items the way I did and not necessarily that they did not know how to handle such questions. If the teachers did not think about classifying the assessment items, then this result would be consistent with the findings of Suah and Ong (2012). They reported that teachers did not prepare a table of specifications to help them determine the cognitive levels of the assessment items they developed.

Although Stiggins et al. (1989) and Delice et al. (2013) found that there was a difference between oral questions and written questions in mathematics in terms of the thinking skills assessed; I found that there was no difference in skills assessment between the oral items and the written items the teachers used in their mathematics classes. Both oral and written items assessed the same low-level thinking skills: recall of factual information, comprehension, and use of routine procedures. The same was true with AP Statistics. John used Group C (items requiring justifying, evaluating, etc; (see Smith et al., 1996) thinking skills more in AP Statistics than he did in mathematics, but Mary did not. Unlike in the study by Senk et al. (1997), where they reported that only geometry teachers used questions that required justification, I could not determine from the data in this study the influence of subject matter on the thinking skills assessed. This is consistent with the findings of Duncan and Noonan (2010). They reported that that subject area did not have any influence on the thinking skills assessed.

The use of open-ended assessment items was almost nonexistent. Neither John nor Mary used open-ended items in mathematics. These results are consistent with those of Senk et al. (1997).

DOI: 10.20429/ger.2015.120201

Limitations of the Study

The participants in this study were chosen by using purposeful sampling. They were chosen on the basis that they were teaching both mathematics and AP Statistics. The two participants were also members of an AP Statistics learning community. It might be difficult to extend the findings of this study to other mathematics teachers who also teach statistics as they may possess different characteristics. The findings of this study, however, can still enlighten us on the issues related to assessment practices of mathematics teachers who also teach AP Statistics.

Implications

In this study, I found that the items teachers used for assessment were "low level." As I have indicated, the teachers in the study did not seem to think of the items in terms of the thinking skills required to do them—considering that the teachers used items that measured the same thinking skills. According to the recommendations of *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), mathematics is more than just applying procedures to routine problems: Mathematics also involves making connections, problem solving, communication, and reasoning. The fact that a number of researchers (Sanchez, 2002; Senk et al., 1997; Stiggins et al., 1989) have examined the nature of the items that mathematics teachers use and reached the same conclusion—teachers tend to use low-level items—could mean that teacher education programs are not spending time in addressing the issues involving the characteristics of the items teachers use in their assessments. Suah and Ong (2012) reported that teachers with less experience were unable to develop their own questions because they lacked the necessary skills. Teacher education programs should put more emphasis on teaching teachers how to create good tasks.

Professional developers also ought to spend a lot of time with practicing teachers to help them mainly in the development of tasks that teachers can use in the classrooms. Teachers need the necessary knowledge to enable them to develop or identify tasks that can measure all the thinking skills, instead of just concentrating on the recall skills. Abraugh and Brown (2006) found that engaging high school mathematics teachers to critically examine the mathematics tasks in terms of levels of cognitive demand influenced the way the teachers thought about the mathematics tasks they used. The study further found that some of the teachers changed the patterns of how they selected the tasks to use in the classroom. Sanchez (2002) also reported the effect of involving teachers in assessment projects. In-service teachers ought to be involved in assessment projects. That way they can acquire the necessary skills to develop their own tasks.

Future Research

A study similar to the current study could be conducted on a large scale. The study could include different counties or different school districts. Several other variables could be considered such as using teachers who are initially trained to teach mathematics and statistics. The issue of the difference between mathematics and statistics has been discussed for a long time. It is important to investigate whether the teachers' knowledge of the difference between mathematics and statistics translates into the identification of tasks. For example, are teachers able to distinguish tasks that assess statistical reasoning from those that that assess mathematical reasoning? Can teachers develop their own tasks that will show the difference between items assessing mathematical reasoning and items assessing statistical reasoning?

Conclusion

The purpose of this study was to investigate the assessment practices of mathematics teachers who also taught AP Statistics. In particular, I investigated the thinking skills that

DOI: 10.20429/ger.2015.120201

assessment items used mathematics and AP Statistics classroom assess. Though a number of issues have been discussed in this report; it is obvious that one study alone cannot address many issues related to assessment. The findings of this study serve as a stepping stone to future research studies. Studies involving statistics have always involved college students. Now may be the right time to start thinking seriously about conducting more studies in middle schools and high schools. As noted on page 2, the GAISE project (Franklin et al., 2007) has been in the forefront advocating the improvement of the teaching of statistics in Grades PreK–12.

Assessment is dynamic. As statistics and mathematics content keep on developing, and as also technology is becoming increasingly important in the two subject areas, there is need to make sure that assessment keeps up with those developments. Studying assessment practices of teachers should be at the center of our education. More importantly, studies involving assessment in statistics are critical, as I have indicated earlier that statistics is a relatively new subject in the secondary school.

References

Arbaugh, F., & Brown, C. (2006). Analyzing mathematical tasks: A catalyst for change. *Journal of Mathematics Teacher Education, 8*, 499–536.

American Statistical Association. (2005). *Endorsement.* Retrieved from http://www.amstat.org/education/gaise/ASAEndorse.htm

Bush, W. S., & Greer, A. S. (Eds.). (1999). *Mathematics assessment: A practical handbook for Grades 9–12*. Reston, VA: National Council of Teachers of Mathematics.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3). Retrieved from .www.amstat.org/publications/jse/v10n3/chance.html

College Board. (2007). *AP Statistics course description.* Retrieved from http://www.collegeboard.com

Delice, A., Aydın, E., & and Seda Çevik, K. (2013). Mathematics teachers' use of questions: Is there a change of practice after the curriculum change? *Eurasia Journal of Mathematics, Science & Technology Education, 9* (4), 417–427.

Duncan, C.R., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *The Alberta Journal of Educational Research, 53*(1), 1–21.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A preK–12 curriculum framework.* Alexandria, VA: American Statistical Association.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal. & J. B. Garfield (Eds.), *The assessment challenges in statistics education* (pp. 1–13). Amsterdam, The Netherlands: IOS Press.

DOI: 10.20429/ger.2015.120201

Moore, D. S. (1988, January). Should mathematicians teach statistics? *College Mathematics Journal*, 19, 3–7

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1991). *Professional standards for school mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Quellmalz, E. S. (1985). Developing reasoning skills. In J. R. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 86–105). New York, NY: Freeman.

Sanchez, W. M. B. (2002). Conceptualizing mathematics teachers' use of open-ended assessment items (Doctoral dissertation, University of Georgia, 2001). *Dissertation Abstracts International, 63*A.

Senk, S. L., Beckmann, C. E., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education, 28,* 187–215.

Smith, G., Wood, L., Coupland, M., & Stephenson, B. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *International Journal for Mathematics Education, Science and Technology*, *27*(1), 65–77.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26*(3), 233–246.

Suah, S. L., & Ong, S. L. (2012). Investigating assessment practices of in-service teachers. *International Online Journal of Educational Sciences*, *4*(1), 91–106.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*, 223–265.

DOI: 10.20429/ger.2015.120201

Appendix

EXAMPLES OF MATHEMATICS AND AP STATISTICS ASSESSMENT TASKS FOR EACH TAXONOMY

Where 1: *mathematics problem* and 2: *AP Statistics problem*.

Group A

Factual knowledge and fact systems

1. State the Intermediate Value Theorem.
2. What is an unbiased statistic?

Group B

Information transfer

1. If $f(x) = x^2 + 1$

   I. Find all the zeros of the function $f(x)$.

   II. Find the x-intercepts for the graph of $f(x)$, or explain why they do exist.

2. Peter and Isaac play are playing in a golf tournament. They play repeatedly and their scores vary. Peter's score X has the N (100, 7) distribution, and Isaac's score Y has the N (95, 9) distribution. Is it reasonable to take the variance of the total score to be $\sigma^2_x + \sigma^2_y = 7^2 + 9^2 = 130$? Explain your answer.

Group C

Justifying and interpreting

1. Two students evaluate the expression $-(3-x)^2$

   The first student writes the following:

$$-(3-x)^2 = -[(3-x)(3-x)]$$
$$= (-3-x)(-3-x)$$
$$= (-3)(-3)+(-3)(-x)+(-x)(-3)+(-x)(-x)$$
$$= 9+(-3x)+(-3x)+(-x)^2$$
$$= 9-6x-x^2$$

   The second student writes the following:

$$-(3-x)^2 = -[(3-x)(3-x)]$$
$$= -[(3)(3)+(3)(-x)+(-x)(3)+(-x)(-x)$$
$$= -[9+(-3x)+(-3x)+(x)^2]$$
$$= -[9-6x-x^2]$$
$$= -9+6x+x^2]$$

   Who is right between the two? Find and explain the error(s)

2. A random sample of 1500 teenagers (ages 12-17) was asked whether they watched basketball online; 990 said that they did.

   a. Construct and interpret a 95 % confidence interval for the population *p*.

   b. Suppose that the results of the survey were used to construct separate 95% confidence intervals for boys and girls. Would the margins of error for those two confidence intervals be the same as, larger than, or smaller than that of the interval you constructed in part (a). Justify your answer.

DOI: 10.20429/ger.2015.120201