International Journal of Scientific Research in Education, OCTOBER 2014, Vol. 7(3), 203-215.

International Journal of Scientific Research in Education (IJSRE)
ISSN: 1117-3259

Developing an Assessment of Statistics Teaching Knowledge

Brenna Haines[i]
Wichita State University, USA.
brenna.haines@wichita.edu

Abstract

This study shows the results of the development of an assessment designed to measure secondary-level statistics teaching knowledge. This study addresses a neglected area of research by creating an online, multiple-choice assessment of secondary-level, in-service Advanced Placement (AP) Statistics teaching knowledge. The study focused on the questions (1) Which assessment items will pass a content validation study? (2) Which items will remain in the assessment after item-level analysis? Based on the results of this study, nine content knowledge and five pedagogical content knowledge multiple-choice items remained in the assessment crossing all content and pedagogical domains. On the basis of these findings, the researcher suggests more item development needs to occur in the domain of student learning.

Keywords: Statistics Teaching and Learning, Assessment Development, Item-Level Analysis, Teaching Knowledge, Content Knowledge, Pedagogical Content Knowledge.

Reference to this paper should be made as follows:

Haines, B. (2014). Developing an Assessment of Statistics Teaching Knowledge. *International Journal of Scientific Research in Education, 7*(3), 203-215. Retrieved [DATE] from http://www.ijsre.com

INTRODUCTION

Since the 1990's, statistics and probability concepts have increasingly become an integral part of mathematics standards across the K-12 curriculum (Common Core, 2010; NCTM, 2000). In addition to the increased interest in statistics and probability, The College Board's Advanced Placement (AP) Statistics course has also seen increasing student enrollment. Today, student enrollment in AP Statistics has jumped to nearly seventeen times the number it was when it began in 1998. Increasing student enrollment in AP Statistics necessitates the need for teachers who are knowledgeable in the subject-area. However, recent research shows that while student enrollment and the number of schools offering AP Statistics continues to grow, the number of capable teachers has not kept pace (Franklin, Hartlaub, Peck, Scheaffer, Thiel, & Freier, 2011).

Recent research in statistics education has also shown that statistics can be quite challenging to teach (Mills, 2007), and, in fact, most mathematics teachers are not comfortable teaching statistics (Franklin, 2010). In fact, this idea is not new. A primary topic at the 1988 International Statistical Institute (ISI) Round Table Conference was the lack of adequate preparation, and other deficiencies of teachers who teach statistics. Specifically, the conference discussed how, at the high school level, there were not enough teachers with formal training in statistics and even fewer who had training in the pedagogy of teaching statistics (Hawkins, 1990).

A potential reason for this problem may be that there often exists a lack of adequate preparation among pre-service secondary statistics teachers as well as minimal opportunities for in-service teachers to gain content and pedagogic knowledge (Franklin, 2010). Consequently, despite the fact that research suggests statistics teachers often lack content and pedagogic knowledge, and have few opportunities to gain such knowledge, no one has attempted to measure content and pedagogic knowledge of those currently teaching the subject. Specifically, there does not exist an instrument to measure if a teacher does or does not possess the content-specific knowledge necessary to teach secondary-level statistics. The focus of this study is on the development of a teaching knowledge assessment specific to AP Statistics due to the popularity of the course.

Purpose

The goal of this study was to develop, validate, pilot and revise a set of items in order to create an Advanced Placement Statistics Teaching Knowledge (APSTK) online, multiple-choice assessment designed to measure content knowledge and pedagogical content knowledge of secondary-level, in-service AP Statistics teachers. Recent research in statistics teaching and learning specifically calls for further examination of the pedagogic knowledge necessary for teaching secondary-level statistics (Casey, 2010; Chance & Rossman, 2006; Groth & Bergner, 2005; Franklin, et al., 2011; Hassad, 2006; Leavy 2010; Liu & Thompson, 2007; Phannkuch, 2006) as well as how teaching knowledge affects subject-specific student learning (Ball, 2008; Ball et al., 2008; Franklin et al., 2005; Hubbard, 1997; Schoenfeld & Kilpatrick, 2008). Development of an APSTK assessment will contribute to these research efforts.

METHODS

General steps for the APSTK assessment construction were adapted from the classical test theory processes outlined in the Crocker and Algina (2008) text entitled *Introduction to Classical and Modern Test Theory*. These steps included identifying a purpose for the assessment, identifying behaviors that represent constructs, writing an initial pool of assessment items, having items reviewed and revised and, finally, piloting the items.

Purpose of the Assessment

Items were designed, piloted and revised in order to create a multiple-choice assessment instrument to measure Advanced Placement Statistics Teaching Knowledge (APSTK) among current, in-service, secondary-level AP Statistics teachers. Content knowledge for teaching AP Statistics (APSCK) and pedagogical content knowledge for teaching AP Statistics (APSPCK) items were developed to represent the over-arching construct of APSTK or total teaching knowledge required for AP Statistics.

APSCK and APSPCK items were written to reflect current AP Statistics curriculum while keeping in mind the work of previous teaching knowledge theorists (Shulman, 1986; Ball et al., 2008; Schoenfeld & Kilpatrick, 2008) as well as established statistics student learning frameworks, such as The Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Franklin & Garfield, 2006) recommendations.

Identifying Teacher Behaviors

Initially, a set of construct behaviors and behavior characteristics were developed to begin to conceptualize the individual components of each of the APSCK and APSPCK constructs. Therefore, a list of behaviors that represent APSCK and APSPCK were compiled. Crocker and Algina (2008) recommend using techniques such as "expert judgment" (p. 68) to verify that specific assessment items do, in fact, make-up the constructs being measured. They recommend that this input be obtained from one or more individuals who have experience with the construct. For this study, the researcher spent six years teaching AP Statistics, so her knowledge was used to outline teacher behaviors that represent APSCK and APSPCK.

Writing an Initial Pool of Assessment Items

According to the American Association for the Advancement of Science (AAAS) Project 2061 Symposium on Assessment National Association for Research on Science Teaching (NARST) Annual Conference (DeBoer, Abell, & Gogos, 2007) item comprehensibility, test-wiseness, and appropriateness of task context issues are some of the most common threats to construct validity. Specifically, test items may contain construct irrelevant features that make it difficult to rely on the answers that examinees choose. In general, teacher respondents should choose the correct answer when they know the idea, and they should choose an incorrect answer when they do not know the idea.

Furthermore, the NARST annual conference (2007) concluded that in order to make valid interpretations of what a teacher knows, a test item must be understandable. Therefore, to ensure that valid interpretations of teacher knowledge it must first be clear what question is being asked. For example, the item must use familiar general vocabulary that is clearly defined and should not use unnecessarily complex sentence structure or ambiguous punctuation that makes the question difficult to comprehend. Finally, the item cannot include words and phrases that have unclear, confusing, or ambiguous meanings. APSCK and APSPCK items were developed with these principles in mind.

In addition, consideration was given to ensure that items include accurate information (including information on diagrams and data tables), and therefore not confusing to teachers who have a correct understanding of AP Statistics. For example, respondents should not have to spend time wondering if there is a mistake in the wording of an item.

As items were developed, consideration was also given to the appropriateness of the context. According to the American Association for the Advancement of Sciences (AAAS) Project 2061 (DeBoer, et al., 2007), in order to make valid interpretations of what teachers know any item that involves a situation with people, objects or events in the real world or otherwise must be checked to be understandable, interesting, and sensible to potential respondents.

To make valid interpretations of APSCK and APSPCK, items were written so that they are not answerable by using test-taking strategies. According to the AAAS Project 2061 (DeBoer, et al., 2007), test-wiseness addresses characteristics of the item that might allow respondents to make a satisfactory response using only general test-taking skills without understanding the idea being tested or mislead respondents into choosing an incorrect answer. For example, in multiple-choice items, one answer (especially the correct answer) should not be distinctly different from the other answers. Finally, items were checked so that qualifying terms such as "always," "never," or "everyone" were avoided; as well as logical opposites, or other obvious distractors that may make it easy to eliminate answers.

According to Crocker and Algina (2008), there should be a balance of items so that the different constructs, or components of a single construct, are represented in proportion to their importance. Therefore, the multiple-choice APSCK items were balanced with an equal number per sub-domain (exploring data, sampling and experimentation, anticipating patterns and statistical inference). In addition, a set of measurable characteristics (problem-solving knowledge, organizing content and student learning) for APSCK also guided and balanced the development of those items.

Item Review

According to Crocker and Algina (2008), one must conduct a content validation study to assess whether items adequately represent a construct of particular interest. In order to ensure valid content for this study, a panel of independent experts was asked to judge whether proposed assessment items represent APSCK or APSPCK. Crocker and Algina (2008) suggest the following four steps to be undertaken during the construction of a content validation study:

1. Define the behavior or performance domain of interest
2. Select a panel of qualified experts in the content area
3. Provide a structured framework for the process of item alignment
4. Collect and summarize the content alignment feedback from the experts

To this end, two experienced AP Statistics teachers as well as two professors in mathematics education at different local universities vetted the initial items designed to assess APSCK and APSPCK. Items were vetted with content alignment criteria set forth by the American Association for the Advancement of Sciences (AAAS) Project 2061 (DeBoer, et al., 2007) to ensure construct validity in terms of comprehensibility, appropriateness of task context and test-wiseness. Appendix A outlines the content alignment criteria questions that were provided to reviewers for each item.

Each reviewer analyzed thirty potential APSCK items and eighteen potential APSPCK items. Each item was deemed acceptable if at least three out of four reviewers identified it as passing seventy-five percent of the alignment criteria questions. Items that were deemed acceptable by three out of four reviewers on less than seventy-five percent of the alignment criteria questions were revised. Items that were not deemed acceptable by three out of four reviewers were not included in further item review.

Sample

The APSTK assessment was piloted to a systematic, random sample of five hundred potential respondents by inviting them to participate online and agree to answer an initial set of twelve multiple-choice content knowledge items and nine constructed-response pedagogical content knowledge items. According to Crocker and Algina (2008) item level analysis can be conducted with relative stability for samples of at least two hundred respondents. Therefore, it was anticipated that initial sampling would result in at least two hundred complete assessments.

Data Collection

APSCK items

APSCK items were presented in multiple-choice format. Additional item writing guidelines for APSCK multiple-choice items followed rules set forth by Halaydna and Downing (1989) and Frey (2005) regarding general item construction, content and answer options. For example item construction rules such as, “use a best answer format”, “minimize examinee reading time” and “avoid trick items” were used during item construction. For item content, rules such as “base each item on an educational or instructional objective”, “use multiple-choice to measure higher level thinking”, “and keep items independent of one another” were also used (Halaydna & Downing, 1989). For answer options, rules such as “answer options should only include one correct answer”, “answer options should be homogenous”, “there should be three to five answer options”, and “negative wording should not be used” were also considered (Frey, 2005).

Crocker and Algina (2008) suggest there should be a balance of items so that different constructs, or

components of a single construct, are represented in proportion to their importance. Therefore, the multiple-choice APSCK items were balanced with an equal number per sub-domain (exploring data, sampling and experimentation, anticipating patterns and statistical inference). In addition, a set of measurable characteristics (problem-solving knowledge, organizing content and student learning) for APSCK also guided the development of those items. For sub-domain content areas and measurable characteristics of APSCK items, please see Table I.

APSCK items were scored dichotomously and examined using item-level analysis. Respondents were given a "correct" rating if they chose the best possible answer or an "incorrect" rating if they chose an answer other than the best answer. A numeric score of "1" per item was given to correct answers, and a "0" was given for incorrect answers.

APSCK assessment items were examined for both item level difficulty and discrimination to determine if an item should be kept or eliminated. An item's difficulty level was defined as being the proportion of examinees that answered the item correctly, and produced a value in a range from 0.00 to 1.00. So, the easier an item was to answer (determined by respondents' ability to choose the correct answer) the higher the item's difficulty coefficient was. For example, an item answered correctly by 85.00% of respondents had a *p*-value of 0.85, while an item answered correctly by 50.00% of respondents had a lower *p*-value of 0.50. An item's difficulty level controls the item's variance (σ^2) because variance $\sigma^2 = p_i q_i$ where p_j = (number of persons with a score of 1 on item j)/ N and $q_j = (1-p_j)$.

In addition, it was assumed that there existed a constant degree of correlation among items, therefore total assessment score variance was considered to be maximized when item difficulty is p_i=0.50. According to Crocker and Algina (2008) item difficulties (or *p*-values) for score interpretation usually fall into a range of 0.60 to 0.80 (p. 312). For this study, an acceptable difficulty level for an item was set at a range of 0.40 to 0.70. Therefore, items falling below or above this range did not pass the item difficulty analysis and were dropped from further analysis.

For those APSCK items that passed the item difficulty analysis, an item discrimination value was calculated using a criterion-based index of discrimination (D). Crocker and Algina (2008) state that using a D-value is recommended because it is easy to calculate and its' interpretation is straightforward. Because content knowledge items were designed to measure the content knowledge construct, a discrimination parameter was calculated to examine how effectively an item discriminates between respondents who have a relatively high level of content knowledge from those with a relatively low level.

For this purpose, an initial APSCK assessment score was established for each individual respondent. The purpose of calculating an individual APSCK total score was meant to establish a criterion score distribution (or cut scores) to determine the item level discrimination parameter (D). Therefore, examinees were split into two groups: those who scored below and above the cut scores. Again, these cut scores were used for item-analysis purposes only, and not meant to imply an average level or range of teaching knowledge among respondents.

Determining the upper and lower bound or cut score has been done several ways in past research. For example, Kelly (1939) found a stable index discrimination index to include the upper 27.00% and the lower 27.00% of the examinees total assessment scores (in Crocker & Algina, 2008). In another analysis, Beuchart and Mendoza (1979) and Englehart (1965) found that when the sample size is relatively large the same results can be obtained by having an upper and lower bound of 30.00% and 50.00%. For the APSTK assessment, the cut score was set at 30.00% for the upper group and 50.00% for the lower group. So, for the purpose of item-analysis only, those respondents who scored in the upper 30.00% possess an adequate level of APSCK and those that scored in the lower 50.00% do not.

Once the upper and lower groups were determined the index of discrimination is calculated as $D = p_u - p_i$ where p_u was the proportion in the upper group who answered the item correctly and p_i was the proportion in the lower group who answered the item correctly. The value of D ranged from -1.00 to 1.00 where positive values indicated that the item discriminates in favor of the upper group and negative values

indicated that the item favors the lower scoring group. Specifically, the following guidelines, Ebel as quoted in Crocker and Algina (2008) will be used:

If $D \geq 0.40$ the item if functioning satisfactorily
If $0.30 \leq D \geq 0.39$, little or no revision required
If $0.20 \leq D \geq 0.29$ the item is marginal and needs revision
If $D \leq 0.19$ the item should be eliminated or completely revised

APSPCK Items

As with the APSCK items, APSPCK assessment items were developed to reflect a measurable characteristic that matches a particular AP Statistics content area. In addition, an attempt was be made to create an equal number of items per sub-domain and measurable characteristic so that each category and definitional component was adequately represented. For sub-domain content areas and measurable characteristics of APSPCK items, please see Table 2.

In addition, a Likert-scale was used to score initial APSPCK constructed response items. A general scoring rubric for these items can be found in Appendix B; however, each APSPCK item had a unique corresponding scoring rubric assigned to it. The APSPCK item scoring rubric format was adapted from the scoring rubric used to grade free-response questions on the AP Statistics exam from 2012. This kept the assessment of APSPCK items closely aligned with student expectations on the AP Statistics exam. Specifically, The College Board (2012) Scoring Guidelines suggest that "essentially correct", "partially correct" or "incorrect" response ratings depend on the particular question being asked. The same held true for APSPCK responses in the current study. For example, if an item had three sections and all three sections were answered correctly, the grade given was a "4" for a complete response.

Inter-rater reliability

To increase reliability, two raters analyzed APSPCK teacher responses using item-specific rubrics. In order to measure the level of agreement among raters, a Cohen's Kappa (κ) coefficient (Cohen, 1960) was calculated and analyzed. Cohen's Kappa was defined as:

$$\kappa = \frac{Pr\,(a) - Pr\,(e)}{1 - Pr\,(e)}$$

where Pr (*a*) was defined as the observed percentage of agreement and Pr (*e*) was defined as the expected percentage of agreement. In general, a Cohen's Kappa value of 1.00 implies perfect agreement and -1.00 implies imperfect agreement. For the current study, a value of 0.70 was considered a satisfactory percentage of agreement among raters.

RESULTS

Expert Panel Review

The initial pool of APSTK assessment items consisted of thirty multiple-choice APSCK items and nineteen constructed-response APSPCK questions. After the expert panel review, twelve APSCK items and nine APSPCK items remained. Table 1 provides specific details regarding the number of APSCK items per content sub-domain and measurable characteristic that remained after the expert panel review.

Table 1: Specification Table for APSCK Items (Listed by Item Number)

AP Statistics Content Sub-Domains	APSCK Measurable Characteristics		
	Problem Solving Knowledge	Organizing Content	Student Learning
Exploring Data	4	7	5
Sampling and Experimentation	1	6	10
Anticipating Patterns	2	11	9
Statistical Inference	3	12	8

In addition, nine constructed-response APSPCK items remained after the expert panel review. Table 2 provides specific details regarding the number of APSPCK items per content sub-domain and measurable characteristic that remained in the assessment.

Table 2: Specification table for APSPCK Items (Listed by Item Number)

AP Statistics Content Sub-Domains	APSPCK Measurable Characteristics		
	Adjust curriculum, instruction and assessment strategies	Recognize and interpret errors and misconceptions	Student thinking and learning
Exploring Data		17	21
Sampling and Experimentation	19	16	
Anticipating Patterns	13	14	20
Statistical Inference	15		18

Five hundred current AP Statistics teachers across the United States were invited to respond to the APSTK online assessment, but only 465 participant invitations reached the intended respondents due to incorrect or invalid contact information. This sampling resulted in 202 partial and 108 complete assessment responses. Response rates were approximately 30.00% for the multiple-choice APSCK items and 10.00% for the constructed-response APSPCK items.

APSCK Item Analysis

Difficulty coefficients and discrimination parameters for APSCK items were calculated by hand and analyzed. For comparison purposes, jMetrik software (Meyer, 2009) was also used to calculate item-level difficulty and discrimination parameters.

Difficulty and discrimination coefficients for the APSCK items are listed in the following table. In some cases, there was a noticeable discrepancy between the hand-calculated coefficients and the coefficients calculated by the jMetrik software. This was due to the handling of missing or incomplete data. Specifically, the hand-calculation method ignored missing data while the jMetrik software treated missing data as having a zero value. Table 3 details the hand-calculated and jMetrik coefficients for the twelve APSCK items in the assessment.

Table 3: Difficulty and Discrimination Coefficients

APSCK Item	Difficulty*	Discrimination*	jMetrik Difficulty	jMetrik Discrimination
10	88	56	84	-0.8
11	85	45	4.4	3
12	80	64	57	26
13	42	39	41	41
14	95	-6	0	29
15	67	50	38	0
16	95	50	0	8
17	41	29	21	49
18	34	14	32	50
19	94	-4	0	12
20	95	13	41	60
21	40	36	14	14

*Hand Calculated

Items were then examined for a final time and those with high difficulty parameters, and or zero or close to zero discrimination parameters were considered for deletion or revision. Ultimately, an item was deleted if it had a negative discrimination parameter from either source, because this meant that the probability of choosing a correct response for that item decreased as a respondent's ability level increased. This means that the item was either poorly written or improperly measuring high-ability respondents.

Specifically, four APSCK items (numbers 10, 11, 15, and 21) were revised and three items (numbers 14, 18, and 19) were deleted. This resulted in nine APSCK items remaining in the assessment after item-level analysis. Table 4 outlines the categories of APSCK items listed by item number on the assessment. It should be noted that only one APSCK item pertaining to student learning remained in the assessment after the pilot data analysis.

Table 4: Final Categories of APSCK Items

	APSCK Measurable Characteristics		
AP Statistics Content Sub-Domains	Problem Solving Knowledge	Organizing Content	Student Learning
Sampling and Experimentation	6	10	
Anticipating Patterns	7	13	
Statistical Inference	8	14	12
Exploring Data	9	11	

APSPCK Item Analysis

Item-level analysis for the APSPCK began with two different raters using item specific rubrics to rate APSPCK individual constructed responses. Cohen's Kappa (Cohen, 1960) was calculated for each item to determine inter-rater reliability. If there was a discrepancy, individual item ratings were reconciled between the two raters and a final rating was determined. These results are listed below in Table 5.

Table5: APSPCK Cohen's Kappa Coefficients

APSPCK Item Number	N	Cohen's Kappa (K)
1	50	0.87
2	42	0.89
3	27	0.72
4	24	0.81
5	20	0.55
6	17	0.66
7	14	1.00
8	11	0.87
9	9	0.44

Cohen's Kappa indicated that raters had high consistency with scoring. Those items with less than the established acceptable level of 0.70 (item numbers 5, 6 and 9) were negotiated a second time among raters.

In order to develop a multiple-choice assessment of both APSCK and APSPCK, constructed-response APSPCK items were changed into a multiple-choice format. This was accomplished for each item by analyzing respondent constructed answers, in combination with item-specific rubrics. For each APSPCK item, a workable set of multiple-choice responses were created using respondent data and placed on a continuum of one to four corresponding to increasing levels of pedagogical content knowledge.

At this point, each APSPCK multiple-choice item had four answer choices with each answer choice signifying a specific level of APSPCK. Individual response data given a "4" rating, for example, were combined and synthesized to make-up a level four answer choice. This level-four answer choice signified the highest level of APSPCK possible for that item. Those responses given a "3" grade were synthesized to produce a level-three answer choice and so on.

At this point, a total of five multiple-choice APSPCK items remained in the assessment. Specifically, the last four original APSPCK items were deleted from further analysis due to low response rate and missing data. The characteristic categories of APSPCK items one through five did not change after data collection and item level analysis. Table 6 outlines the categories of APSPCK items.

Table 6: Final Categories of APSPCK Items

	APSPCK Measurable Characteristics		
AP Statistics Content Sub-Domains	Adjust curriculum, instruction and assessment strategies	Recognize and interpret errors and misconceptions	Student thinking and learning
Exploring Data		5	
Sampling and Experimentation			4
Anticipating Patterns	1	2	
Statistical Inference	3		

DISCUSSION/CONCLUSION

Assessment items were created to measure teaching knowledge for the College Board's AP Statistics course. To date, there does not exist a description or measure of the amount or types of teaching knowledge for secondary-level statistics in general, or AP Statistics specifically. Using detailed item development and analysis techniques, an APSTK assessment consisting of multiple-choice items was created to reflect constructs pertaining to AP Statistics content knowledge (APSCK) and AP Statistics pedagogical content knowledge (APSCK).

From a total of forty-eight potential content knowledge and pedagogical content knowledge items, an expert panel review approved twenty-one items for further consideration (twelve APSCK and nine APSPCK). After piloting the assessment and item-level analysis, the APSTK assessment was left with nine APSCK items and five APSPCK items.

Specifically, there were two APSCK items left in each content sub-domain of Sampling and Experimentation, Anticipating Patterns and Exploring Data, as well as three items left in the Statistical Inference content sub-domain. Furthermore, four items of the remaining nine were in the measurable content knowledge characteristics category of Problem Solving Knowledge, as well as the Organizing Content category. Finally, only one item remained in the Student Learning content knowledge characteristic category.

Item-level analysis also produced a set of multiple-choice APSPCK items. Due to incomplete or missing data, only the first five APSPCK items were analyzed and remained in the APSTK assessment after data collection. Respondent data was combined, synthesized and rated using item-specific rubrics to produce a set of multiple choice answers tied to specific levels of increasing pedagogical content knowledge. This process resulted in each APSPCK item having four answer choices reflecting varying amounts of pedagogical content knowledge.

This analysis resulted in one remaining APSPCK item for each content sub-domain of Exploring Data, Sampling and Experimentation, and Statistical Inference. However, two pedagogical content knowledge items remained in the content sub-domain of Anticipating Patterns. In addition, two APSPCK items were left per measurable characteristic of Adjusting Curriculum, Instruction and Assessment Strategies and Identifying Errors and Misconceptions. However, only one APSPCK item remained for the measurable characteristic entitled Student Thinking and Learning.

It is recommended that future research include more item development of both APSCK and APSPCK items. Specifically, APSCK items that had either a high difficulty coefficient or low (or negative) discrimination coefficient were considered poorly written, and need further attention. In addition, the measurable characteristic category of Student Thinking and Learning was under-represented by remaining APSCK and APSPCK items. Therefore, more items targeting this specific area need to be developed.

REFERENCES

Ball, D. L., Hoover, M., & Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education, 59*(1), 389-407.

Casey, S. A. (2010). Subject Matter Knowledge for Teaching Statistical Association. *Statistics Education Research Journal*, *9*(2), 50-68.

Chance, B., & Rossman, A. (2006). *Investigating statistical concepts, applications, and methods*. Belmont, CA: Duxbury.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 1,* 37-46.

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory.* Mason, OH: Cengage Learning.

Assessment Linked to Science Learning Goals: Probing Student Thinking During Item Development. Paper presented at the AAAS Project 2061 Symposium on Assessment National Association for Research on Science Teaching Annual Conference, New Orleans.

Franklin, C., Hartlaub, B., Peck, R., Scheaffer, R., Thiel, D., & Freier, K. T., (2011). AP Statistics: Building Bridges Between High School and College Statistics Education. *The American Statistician*, *65*(3), 177-182.

Franklin, C. (2010). Are Our Teachers Prepared to Provide Instruction in Statistics at the K–12 Levels? *National Council of Teachers of Mathematics.* Retrieved from http://www.nctm.org/resources/content.aspx?id=1776

Franklin, C., & Garfield, J. (2006). The Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project: Developing Statistics Education Guidelines for Pre K-12 and College Courses. In G. F. Burrill, (Ed.), *Thinking and Reasoning about Data and Chance: Sixty-eighth NCTM Yearbook* (pp. 345-375). Reston, VA: National Council of Teachers of Mathematics.

Frey, B., (2005). Item-Writing Rules: Collective Wisdom. *Teaching and Teacher Education, 21*(4), 357-364.

Groth, R. E., & Bergner, J. A. (2005). Pre-Service Elementary Teachers' Metaphors For The Concept Of Statistical Sample. *Statistics Education Research Journal*, *4*(2), 27-42.

Haladyna, T. M., & Downing, S. M. (1989). A Taxonomy of Multiple-Choice Item Writing Rules. *Applied Measurement in Education, 2*(1), 37-50.

Hassad, R. A. (2006, July). *Reflection On Training, Experience, and Introductory Statistics A Mini-Survey Of Tertiary Level Statistics Instructors*. Paper presented at the International Association for Statistical Education (IASE) International Conference on the Teaching of Statistics, Salvador, Bahia, Brazil.

Hawkins, A. (1990). Training Teachers to Teach Statistics. *Teaching Statistics, 11*(3), 83-84.

Leavy, A. M. (2010). The Challenge Of Preparing PreService Teachers To Teach Informal Inferential Reasoning. *Statistics Education Research Journal*, *9*(1), 46-67.

Liu, Y., & Thompson, P. (2007). Teachers' Understanding of Probability. *Cognition and Instruction*, *25*(2), 113-160.

Meyer, J. P. (2009). jMetrik (computer software). Virginia: University of Virginia.

Mills, J.D. (2007). Teacher Perceptions and Attitudes About Teaching Statistics in P-12 Education. *Educational Research Quarterly*, *30*(4), 16-34.

Pfannkuch, M. (2006). Comparing Box Plots Distributions: A Teacher's Reasoning. *Statistics Education Research Journal*, *5*(2), 27-45.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15*(2), 4-14.

Schoenfeld, A. H., & Kilpatrick, J. (2008). Toward a Theory of Proficiency in Teaching Mathematics. In Wood, T. & Tirosh, D. (Eds.), *International Handbook of Mathematics Teacher Education: Vol. 2. Tools and Processes in Mathematics Teacher Education,* (pp. 1-35). The Netherlands: Sense Publishing.

The College Board (2012). *Advanced Placement Statistics Free-Response Questions and Scoring Guidelines*. Retrieved from: https://apstudent.collegeboard.org/apcourse/ap-statistics/exam-practice

---

[i] Dr. Brenna Haines is an Assistant Professor of Mathematics Education in the College of Education at Wichita State University in Wichita, Kansas. Her areas of specialization and research include secondary-level statistics teaching and learning, pedagogical content knowledge for teaching statistics, and measurement and assessment techniques

Appendix A

Content Alignment Criteria Questions (American Association for the Advancement of Sciences (AAAS) Project 2061)

- Is it clear what the item is asking?
- Are mathematical and statistical representations accurately stated?
- Does the item use familiar general vocabulary (not mathematical terminology)?
- Does the item use unnecessarily complex sentence structure or ambiguous punctuation that makes the question confusing?
- Does the item use words and phrases that have unclear, confusing, or ambiguous meanings?
- Does the item include inaccurate information that may be confusing to teachers who have a correct understanding of the subject?
- If present, are diagrams, graphs, and data tables clear and comprehensible?
- Is the context of the item familiar to AP Statistics teachers?
- Is the context of the item easy to understand so that respondents do not have to spend a lot of time trying to figure out what the context means?
- Is the information and quantities used reasonable and believable?

Appendix B

General APSPCK Item Scoring Rubric (adapted from The College Board, 2012)

Score/Description	Ratings
4 / Complete Response	All three parts* essentially correct
3 / Substantial Response	Two parts essentially correct and one part partially correct

2 / Developing Response	Two parts essentially correct and one part incorrect OR One part essentially correct and two parts partially correct OR One part essentially correct and one part partially correct OR All three parts partially correct
1 / Minimal Response	One part essentially correct and two parts incorrect OR Two parts partially correct and one part incorrect

* Assuming a three-part question