http://dx.doi.org/10.7468/jksmed.2013.17.4.243
J. Korean Soc. Math. Educ., Ser. D, Res. Math. Educ.
Vol. 17, No. 4, December 2013, 243–266

Designing an Assessment to Measure Students' Inferential Reasoning in Statistics: The First Study, Development of a Test Blueprint

PARK, Jiyoon
Federation of State Boards, 124 West Street South, Alexandria, VA 22314, USA;
Email: parkx666@umn.edu

(Received May 21, 2013; Revised December 27, 2013; Accepted December 27, 2013)

Accompanied with ongoing calls for reform in statistics curriculum, mathematics and statistics teachers purposefully have been reconsidering the curriculum and the content taught in statistics classes. Changes made are centered around statistical inference since teachers recognize that students struggle with understanding the ideas and concepts used in statistical reasoning. Despite the efforts to change the curriculum, studies are sparse on the topic of characterizing student learning and understanding of statistical inference. Moreover, there are no tools to evaluate students' statistical reasoning in a coherent way. In response to the need for a research instrument, in a series of research study, the researcher developed a reliable and valid measure to assess students' inferential reasoning in statistics (IRS). This paper describes processes of test blueprint development that has been conducted from review of the literature and expert reviews.

Keywords: statistical inference, inferential reasoning in statistics, informal inference, formal inference
MESC Classification: D65
MSC2010 Classification: 97D60

1. INTRODUCTION

The ability to draw inferences from data is a part of everyday life as people are confronted with situations where they need to critically review data-based claims (Garfield & Ben-Zvi, 2008). Understanding of statistical inference is important in scientific research since the concepts and processes in statistical inference are used in all empirical studies (Sotos, Vanhoof, Van den Noortgate & Onghena, 2007).

Many misunderstandings have been reported that people are confused about the concepts and processes in statistical inference (Falk & Greenbaum, 1995; Haller & Kraus, 2002; Wilkerson & Olson, 1997; Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007). For example, Tverky and Kahneman (1971) showed that people believe that any sample must be similar to the population, regardless of its sample size. More recently, there have been studies about people's difficulty understanding hypothesis testing. Specifically, research has revealed that students have difficulty understanding—the definition of the hypotheses (Vallecillos & Batanero, 1997), the definition of significance level and the *p*-value (Falk, 1986), and the logic of hypothesis testing (Vallecillos, 1999) when they first learn about those concepts.

In the past few years, statistical educators have looked for new ways to help students build an understanding of statistical inference, in light of current research and new developments in the practice of statistics. As a way to support a coherent understanding of the concepts and processes in statistical inference, Wild, Pfannkuch, Regan & Horton (2011) suggest a learning pathway that introduces some of the "big ideas" behind inference before teaching *formal statistical inference.* Garfield & Ben-Zvi (2008) address that ideas of inference should be introduced informally at the beginning of the course, such as having students become familiar with seeing where a sample corresponds to a distribution of sample statistics, based on a theory or hypothesis.

The big ideas of inference that can be taught before formal inference, suggest two content areas in statistical inference—informal statistical inference (ISI) and formal statistical inference (FSI). In this paper, these terms are used to refer to the content areas of statistical inference. The topics of ISI include: the concept of uncertainty; properties of aggregate data; recognizing sampling variability; the concept of unusualness; (informal) generalization from a sample to a population; (informal) comparison between two samples. The concepts involved in formal statistical testing (e.g., p-value, statistical significance, hypothesis tests, confidence intervals) are categorized as FSI. In addition, the topics of foundations of formal statistical inference (e.g., sample representativeness, sample variability, sampling distribution) are also included in this category given that they are foundational to understanding formal statistical inference (e.g., Chance, delMas & Garfield, 2004).

2. STUDY PURPOSE AND TARGET POPULATION

Now that understanding the basic idea of statistical idea is essential in learning statistics (GAISE: American Statistical Association, 2005) and that there have been pervasive misunderstandings about the concepts, it is important to have an assessment tool to better understand how students interpret ideas of statistical inference. Despite increased interest in informal inferential reasoning and efforts to characterize it, there are no assessments of measuring informal reasoning. Studies are sparse on the topic of how informal inferential reasoning relates to reasoning about formal statistical inference. There are existing instruments used in statistics education research and evaluation to measure students' reasoning in statistics (e.g., The Statistical Reasoning Assessment (SRA): Garfield, 1998; The Statistics Concepts Inventory (SCI): Reed-Rhoads, Murphy & Terry, 2006; and the Assessment Resource Tools for Improving Statistical Thinking (ARTIST), Garfield, delMas & Chance, 2002). Although these instruments assess important outcomes (e.g., assessing students reasoning, thinking, and conceptual understanding), the topics assessed in these instruments do not cover the full domain of reasoning about statistical inference. Moreover, the existing instruments have not been developed or validated using modern measurement models (e.g., item response theory) that provide ample information about properties of test and items (e.g., test validity, item difficulty, item discrimination). Therefore, there is a need for a new instrument that is developed and validated using modern measurement theory so that the results from the assessment provide reliable and valid interpretations. In response to the need for a new research instrument, in a series of study, the researcher describes development processes of an assessment designed to measure students' inferential reasoning in statistics (IRS). As a first step, this paper describes a process of test blueprint development that has been conducted with theoretical and empirical approaches.

Most of the theoretical and empirical data collected in this study is based on education in United States. However, this paper this study can be generalized to a larger population not limited to the students in U.S. given that concepts and content domains in statistical inference are the same across the country. In addition, this assessment is intended to measure *reasoning* as a latent trait or construct rather than to assess students' current *knowledge*; the test for the former case is not necessarily related to specific curriculum.

3. TERMINOLOGIES USED IN THIS STUDY

Literature about statistical inference uses different terms interchangeably (e.g., statistical inference, inferential reasoning in statistics, and reasoning about statistical inference). Specifically, research literature seems to use the two terms without distinguishing between *statistical inference* and *reasoning about statistical inference*. For instance, in Sotos et al. (2007), the researchers use the term statistical inference as a content domain that includes several topics in it. However, in Zieffler, Garfield, delMas & Reading (2008), statistical inference refers to a reasoning process. To clarify the uses of the terms, this study refers to the term statistical inference as a *content domain* that involves the concepts and ideas related to inferential statistics. Also, inferential reasoning in statistics (IRS) is used as reasoning that people uses when drawings conclusions from data.

4. METHODS

4.1. Developing a Test Blueprint from the Literature Review

In a well-designed test blueprint, it is ensured that there is a sound relationship between the test contents in the blueprint and the construct the proposed test is intended to measure. Then, the test blueprint itself provides evidence based on the test content when it represents the content domain (AERA, APA & NCME, 1999). In order to make an agreement on the test score interpretation and uses, it is required to decide on the scope of domains that will be covered in the assessment. However, since there is no criterion reference of IRS, the literature of informal and formal statistical inference was reviewed first. After the content domains were chosen, the types of reasoning to be assessed in the domains were specified based on what the previous researchers considered as important to be captured, which resulted in a preliminary test blueprint. Misunderstandings and difficulties in statistical interference found in research literature were also categorized.

4.2. Expert Review of the Preliminary Test Blueprint

The preliminary test blueprint was reviewed by content experts, and evaluation reports were gathered to examine the adequacy of the test blueprint as a framework to represent the content domains. According to *Testing Standards*, qualified experts can judge the representativeness of the chosen test contents, and their judgments of the relationship between parts of the test and the construct also provide *evidence based on test content* (AERA et al., 1999). The experts who participated in the review process are described below, along with their credentials. The procedures of how they evaluated the preliminary blueprint follow.

Participants

The preliminary test blueprint developed from the literature was reviewed first by five experts. The two (internal) experts are professionals in the program of statistics education at the University of Minnesota. To recruit external experts in different background (countries, research area focused, etc.), the author contacted eleven potential professionals of statistics educators to ask them to evaluate the test blueprint in early May 2011. These reviewers were selected based on their background and research interests. It was also notable that the pool of reviewers has diversity in terms of their expertise and their level of teaching (*Testing Standards 1.7:* AERA et al., 1999). The email invitation letter and evaluation form were sent out to each of the potential reviewers, and three of them agreed to participate in the review process for the test blueprint. All three reviewers were statistics educators who were actively engaged researchers in the area of statistics education. The first reviewer has published many research studies about students' statistical inference, specifically utilizing technological tools or hands-on activities at the secondary and undergraduate levels in New Zealand.

The second reviewer's expertise is the development of statistics curricula, technological tools, and resources for teaching statistics. He has published in many research journals, specifically about how people elicit and acquire statistical reasoning at work. He is working in the Netherlands.

The third reviewer is an instructor in the Department of Statistics at a college in the Midwest area in the U.S. His expertise is in teaching rather than in research, but he has also been involved in several research projects about the topic of statistical inference. It was expected that his professional experience as a teacher of statistics would provide a valuable perspective in terms of a practical sense of assessing students' inferential reasoning. In addition, he was an introductory statistics textbook author who designed an innovative curriculum focused on developing IRS.

Procedures

During the entire process of developing a preliminary blueprint, the author had continuous discussions with the internal experts until an agreement was reached for the preliminary blueprint. Thus, only the reviews from the external experts are reported and analyzed in this paper. Feedback on the preliminary test blueprint was collected from the three experts. Each reviewer was provided with a preliminary test blueprint and an evaluation form. The reviewers were asked to provide ratings for their agreement that the test blueprint was adequate as a framework to develop an instrument to assess the IRS in general (See the evaluation form for the questions in Table 1). Specific evaluation questions were also provided, asking the reviewers to rate the degree to which they agreed that the topics and learning goals documented in the blueprint represent the content domain (AERA et al., 1999). The reviewers were also asked to provide suggestions for changes if an item received a rating of less than 2. Items were judged to have a sufficient level of quality if they had a mean rating of 3 (agree) or higher. For items with mean ratings of less than 3, the reviewers' suggestions for the item changes were carefully reviewed and discussed with an internal expert. In addition, the reviewers' comments on the free-response evaluation questions (e.g., whether there was anything missing from the content of the blueprint related to the constructs of informal and formal statistical inference) were also considered in revising the blueprint.

The feedback obtained from the reviewers was prioritized, restricting the topics and learning goals that would be included in the test blueprint. However, several times of individual meetings were held with the internal expert to discuss the reviewers' suggestions. To decide whether or not the suggested changes would be made in the blueprint, several aspects of the blueprint development were considered such as the scope of the domains (statistical inference, ISI and FSI) delineated from the literature review and topics taught in introductory statistics courses in the U.S. As a result, the final version of the test blueprint was produced.

5. RESULTS

5.1. A Test Blueprint Developed from the Literature Review

The initial test blueprint was built from the literature about IRS. Representing the content domains of IRS, the literature was centered on two areas: Informal statistical inference (ISI) and Formal statistical inference (FSI). These two content areas were used as structure of a construct IRS providing the scope of the content to be covered in the assessment.

The definitions of the construct IRS, and two content domains ISI and FSI, were revisited. In this study, ISI was defined as a domain of statistical inference that involves informal processes of making arguments to support inferences about unknown populations based on observed samples not necessarily using standard statistical procedures. FSI was defined as a domain of statistical inference that involves making a conclusion about population from samples or to formally test hypotheses, using standard statistical methods. The topic category of sampling distribution was considered to represent foundations of statistical inference. The topic of hypothesis testing was used as the second category representing the concepts and ideas of formal statistical inference. Therefore, two content areas of FSI were considered as the main topics in this domain—sampling distributions and hypothesis testing. As a result, the domains of the blueprint were categorized into three areas: informal inference (Inf), sampling distribution (SD), and hypothesis testing (HT).

For the topic of sampling distributions, five content domains were culled from the literature: the concepts of samples and sampling; the Law of Large Numbers; population distribution and frequency distribution; population distribution and sampling distribution; and the Central Limit Theorem. The literature review resulted in a preliminary test blueprint, which is shown in Appendix A.

5.2. Expert Review of the Preliminary Test Blueprint

Results of evaluation ratings

Three professionals in statistics education provided their feedback and suggestions on the preliminary test blueprint. Table 1 presents the results of the experts’ ratings for each evaluation question. As shown in the table, the experts generally agreed that the content domains and learning goals listed in the preliminary blueprint represent the target domains of ISI and FSI. It also appeared that the learning goals identified are adequate to assess students' ISI and FSI. However, there are two evaluation questions that one expert assigned to "*disagree*": question 4 and question 8.

Results of the suggestions and comments

In addition to the ratings for the questions to evaluate the adequacy of the contents in the test blueprint, the experts were also requested to identify any important content domains in ISI and FSI not listed in the blueprint. There were common suggestions made from two reviewers. First of all, reviewers 1 and 2 suggested including real world applications in the blueprint. Reviewer 1 commented, "There is no attention to the inferences about the real world or contextual knowledge" in the current version. It was also suggested that the current blueprint had too much focus on the "limited population" in the categories of SD (sampling distribution) and HT (hypothesis testing; Reviewers 1 and 3). One of the reviewers noted, "One can conceptualize a process as an infinite, undefined population." Similarly, another reviewer commented that there is no content from an experimental perspective saying, "It only talks about samples from limited populations."

Another common suggestion was provided about the topic of "effect size" (Reviewers 2 and 3). In the category of HT-2, the topic covers definitions of *P*-value and statistical significance. In addition to the *P*-value, a reviewer suggested to include consideration of "how large is the effect," which is related to the concept of the effect size. A similar comment was made by another reviewer with a suggestion of adding the "data quality or soundness of the method" to the current blueprint.

Specific suggestions were also provided regarding additional topics to be included in the test blueprint. The topics are:

- Correlation and regression (Reviewer 1).
- Using models in ISI (Reviewer 1).
- Using meta-cognitive awareness of what inference is as opposed to performing procedures (Reviewer 1).
- Confidence intervals (Reviewer 2).
- In the category of HT-6, add designing a test to compare two groups in an experiment, not just from populations (Reviewer 2).
- Consider including randomization and bootstrapping methods (Reviewer 2).
- In the category SD-2, include "biased sampling" for sampling representativeness (Reviewer 3).

These suggestions were reviewed carefully by the author, and were also reviewed with an internal advisor. Discussion between the author and internal advisor centered around whether or not these topics should be included. The definition and the domains that the proposed assessment targets were prioritized for the decision (See Appendix C for the details).

Table 1. Evaluation Questions and Ratings Made by Experts

Item	Evaluation Questions	Strongly Agree	Agree	Dis-agree	Strongly Disagree
1	The topics of the blueprint represent the constructs of *informal* statistical inference.	X	XX		
2	The topics of the blueprint represent the constructs of *formal* statistical inference	X	XX		
3	The learning goals of the blueprint are adequate for developing items to assess students' understanding of *informal* statistical inference.	X	XX		
4	The learning goals of the blueprint are adequate for developing items to assess students' understanding of *formal* statistical inference.	X	X	X	
5	The set of learning goals is well supported by the literature.	X	XX		
6	The learning goals are clearly described.		XXX		
7	The categories of the blueprint are well structured.		XXX		
8	The blueprint provides a framework of developing a test to assess informal and formal statistical inference.	X	X	X	

Table 2 summarizes the changes implemented from the reviewers' comments. The rationale for whether those comments were implemented or not appears in Appendix C. There were topics that the reviewers suggested to include that were not implemented in the blueprint. For example, one reviewer suggested adding content about "correlation and regression." However, these were considered as *literacy* or part of descriptive statistics rather than a topic of inferential reasoning. Another reviewer commented that ISI might also include "meta-cognitive awareness", but we decided that the topic of meta-cognition does not fit the definition of ISI. In addition, there was no literature found regarding this topic as part of ISI. The changes made from the expert reviews resulted in the final version of the blueprint (See Appendix B). In the last review process of the blueprint, the acronyms representing the topic categories, SD (sampling distribution) and HT (hypothesis tests), were changed to SampD and Stest, respectively, to avoid confusion: in statistics, the acronym of SD is mostly used to represent standard deviation.

Table 2. Changes to Test Blueprint Implemented from Expert Reviews

Category	Changes Suggested	Changes Made in the Blueprint
Inf	Include real world or contextual knowledge	Added some learning goals to inferential reasoning in a given context
Inf	Include learning goals about "Using models in informal inferential reasoning"	In two categories, informal inference and formal inference, the learning goals of setting up the null model in a given context was added
Inf	Include using meta-cognitive awareness of what inference is as opposed to performing some techniques	Not included in the blueprint
SD and HT	Too focused on the limited population: Add a process as an infinite (undefined) population; Add statistical testing in experiments	Added the topic categories, DE (designs of study) and EV (evaluation of study) to capture students' understanding of the characteristics of different types of studies
HT	Include the learning goals about an understanding of effect size	In a new category of EV, added the learning goal, "Being able to evaluate the results of hypothesis testing considering —sample size, practical significance, effect size, data quality, soundness of the method, etc."
HT	Include data quality, soundness of the method etc.	The topic category, "Evaluation of HT (EV)," was separated out from the Hypothesis Testing categories since this topic is more about assessing how to interpret and evaluate the results from statistical testing by integrating different kinds of information in a given study (e.g., random assignment, sample size, data quality). The learning goal about, "Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, effect size, data quality, soundness of the method, etc.)," was included in this EV category.
SD or HT	Include a topic category on Confidence Intervals	The topic category, "Inference about Confidence Interval, CI" was added.
SD-2	Add a topic of recognizing "biased sampling" for sampling representativeness	The topic of the "Law of Large Numbers" was changed to "sample representativeness" to assess whether students realize the importance of unbiased sampling (quality of samples), in addition to a large sample (sample size)
HT-6	Add designing a test to compare two groups in an experiment	In ST-3 (changed from a category of HT), the learning goal, "designing a statistical test to compare two groups in an experiment," was added.

Category	Changes Suggested	Changes Made in the Blueprint
HT	Include randomization and bootstrapping methods	Not included as a separate learning goal, but will be assessed in a way so that items get at students' reasoning about the ideas involved in randomization and bootstrap methods. Considering that hypothesis testing based on a normal distribution-based approach is not the only way of statistical testing, the original category about hypothesis testing (HT) was changed to statistical testing (ST), which includes randomization or bootstrap methods.
In general	Add the topics, correlation and regression	Not included in the blueprint since the suggested topics were considered as not being in IRS defined in this study.

6. SUMMARY AND NEXT STEP

This study developed a test blueprint as a first step to design an assessment to measure students' inferential reasoning in statistics. The proposed assessment would help mathematics and statistics teachers understand how students interpret the concepts and ideas of statistical, so it gives the teachers useful information. As an assessment to measure a construct (in this study, IRS), it is necessary that the assessment covers multiple aspects of IRS (comprehensiveness of the test content) and that the test blueprint describing topics and learning goals helps instructors know what to look for when assessing IRS (a detailed and clear description of the blueprint).

The test domains were specified based on a thorough literature review, and the test blueprint was developed laying out important topics and learning goals of each domain. To ensure that the domains are representative to the contents in statistical inference and that the topics and learning goals are adequate to be measured in each domain, contents experts who are actively engaged in the field of statistics education were invited to evaluate the blueprint. For the preliminary version of test blueprint created from the literature review, experts agreed that the topics and learning goals are comprehensive and representative as test domains of IRS.

The reviewers generally considered the blueprint as a good resource to be used as a framework in assessing statistical inference. They acknowledged that the test blueprint covered multiple aspects of IRS. Suggestions and comments made from the reviewers were discussed with internal experts to make a decision on whether we implement in the final version of blueprint. The target population, the test domains, and test purposes were prioritized for the decision.

Using this test blueprint, in the next stage of this research study, the researcher will develop an assessment to measure students' inferential reasoning in statistics. This assessment would be different from conventional statistics tests that we can see in the textbooks in that the proposed test will focus on assessing students' ability to reason not simply asking memorized concepts or procedural knowledge. Moreover, this assessment will provide statistics instructors with useful information of what students misunderstand and how to design the courses by evaluating students' statistical reasoning in a well-structured, coherent content domain. Thus, instructors can get better ideas of designing statistics curriculum to help students develop their inferential reasoning in statistics

REFERENCES

AERA; APA & NCME (1999). *Standards for educational psychological testing*. Washington, DC: AERA.

American Statistical Association (2005). *GAISE College Report.* Retrieved from ASA GAISE College Report Web site: http://www.amstat.org/education/gaise/GAISECollege.htm

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning* **2(1–2)**, 75–97. ME **2001f.**04627

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review* **48(3)**, 378–399. Available from: http://scholasticadministrator.typepad.com/thisweekineducation/files/the_case_against_statistical_significance_testing.pdf

Chance, B.; delMas, R. & Garfield, J. (2004). Reasoning about sampling distributions. In: D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295–323). Dordrecht, Netherlands: Kluwer Academic.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, **49(12)**, 997–1003.

delMas, R. C.; Garfield, J. B. & Chance, B. L. (1999). A model of classroom research in action: developing simulation activities to improve students' statistical reasoning. *J. Stat. Educ.* **7(3)**, 80K. ME **2000b.**01311 Available from: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

Falk, R. (1986). Misconceptions of statistical significance. *J. Struct. Learn.* **9(1)**, 83–96. ME **1986h.**02861

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology* **5(1)**, 75–98.

Garfield, J. (1998). The Statistical Reasoning Assessment: Development and Validation of a Research Tool. In L. Pereira Mendoza (Ed.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 781–786). Voorburg, The Netherlands: International Statistical Institute.

Garfield, J. & Ben-Zvi, D. (2008). *Developing Students Statistical Reasoning: Connecting Re-*

search and Teaching Practice. Dordrecht, Netherlands: Springer. ME **2009b.**00447

Garfield, J., delMas, R., & Chance, B. (2002). "The Web-based ARTIST: Assessment Resource Tools for Improving Statistical Thinking" Project *J. Stat. Educ.* **7(3)**. [Online]. www.amstat.org/publications/jse/v5n3/giraud.html

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research* **7(1)**, 1–20.

Lipson, A. (2003). The role of the sampling distribution in understanding statistical inference. *Mathematical Educational Research Journal* **15(3)**, 270–287. Retrieved from: http://files.eric.ed.gov/fulltext/EJ776331.pdf

Liu, Y. & Thompson, P. (2009). Mathematics teachers' understandings of proto-hypothesis testing. *Pedagogies* **4(2)**, 129–138.

Makar, K. & Rubin, A. (2009). A framework for thinking about informal statistical inference. *SERJ - Stat. Educ. Res. J*. **8(1)**, 82–105. ME **2009e.**00573 Retrieved from: http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf

Metz, K. E. (1999). Why sampling works or why it can't: Ideas of young children engaged in research of their own design. In: R. Hitt and M. Santos (Eds.), *Proceedings of the Twenty-First Annual Meeting of the North American Chapter of the International Group for the Psychology of Education* (pp. 492–498). Columbus, OH: 1999 ERIC Clearinghouse of Science, Mathematics, and Environmental Education.

Mittag, K. C. & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher* **29(4)**, 14–20.

Mokros, J. & Russell, S. J. (1995). Children's concepts of average and representativeness. *J. Res. Mathe. Educ.* **26(1)**, 20–39. ME **1995f.**03881

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* **5(2)**, 241–301.

Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In: G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267–294). Dordrecht, Netherlands: Kluwer Academic.

Reed-Rhoads, T.; Murphy, T. J. & Terry, R. (2006). The Statistics Concept Inventory: An Instrument for Assessing Student Understanding of Statistics Concepts, SIGMAA on Statistics Education session *First Steps for Implementing the Recommendations of the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*, Joint Mathematics Meetings, San Antonio, January 2006.

Rosenthal, R. (1993). Cumulating evidence. In: G. Keren (Ed.), *A handbook of data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Rubin, A.; Bruce, B. & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 314–319). Dunedin, New Zealand: International Statistical Institute.

Rubin, A.; Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In: A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS) held at Salvador, Bahai, Brazil, July 2–7, 2006* (CD-ROM). Voorburg, Netherlands: International Statistical Institute.

Saldanha, L. (2004). *"Is this sample unusual?": An investigation of students exploring connections between sampling distributions and statistical inference*. Unpublished Ph.D. Thesis. Nashville, TN: Vanderbilt University.

Saldanha, L. & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educ. Stud. Math.* **51(3)**, 257–270. ME **2003d.**03486

Schwartz, D. L.; Goldman, S. R.; Vye, N. J.; Barron, B. J. & Cognition Technology Group at Vanderbilt (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In: S. Lajoie (Ed.), *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K–12* (pp. 233–273). Hillsdale, NJ: Erlbaum.

Sedlemeier, P. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavior Decision Making* **10**, 33–51.

Sotos, A. E. C.; Vanhoof, S.; Van den Noortgate, W. & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review* **2**, 98–113.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* **76**, 105–110. [Reprinted in: D. Kahneman, P. Slovic & A. Tversky (1982), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.]

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* **18**5**(4157)**, 1124–1131. Retrieved from:
http://files.eric.ed.gov/fulltext/EJ776331.pdf

Vallecillos, A. (1999). Some empirical evidences on learning difficulties about testing hypotheses. In: *Proceedings of the 52 session of the International Statistical Institute* (pp. 201–204). Helsinki: International Statistical Institute. Retrieved from:
https://www.stat.auckland.ac.nz/~iase/publications/5/vall0682.pdf

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypotheses testing by university students. *Themes in Education* **3(2)**, 183–198.

Vallecillos, A. & Batanero, C. (1997). Activated concepts in the statistical hypothesis contrast and their understanding by unversity students. *Reserchers en Didactique des Mathematiques* **17**, 29–48.

Vanhoof, S.; Sotos, A.; Onghena, P. & Verschaffel, L. (2007). Students' reasoning about sampling distribution before and after the sampling distribution activity. In: *Proceedings of the 56 session of the International Statistical Institute*, Lisbon, Spain, International Statistical Institute.
[Online]: www.stat.auckland.ac.nz/~iase/publications/isi56/CPM80_Vanhoof.pdf.

Wagner, D. A. & Gal, I. (1991). *Project STARC: Acquisition of statistical reasoning in children*.

(Annual Report: Year 1, NSF Grant No. MDR90-50006). Philadelphia, PA: Literacy Research Center, University of Pennsylvania.

Watson, J. M. & Moritz, J. B. (2000a). Developing concepts of sampling. *J. Res. Math. Educ.* **31(1)**, 44–70. ME **2000e.**03613

_____ (2000b). The longitudinal development of understanding of average. *Math. Think. Learn.* **2(1–2)**, 11–50. ME **2000d.**02870 ME **2001f.**05345

Well, A.; Pollastek, A. & Boyce, S. (1990). Understanding of the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes* **47**, 289–312.

Wild, C. K.; Pfannkuch, M.; Regan, M. & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. J. *Royal Statistical Society* A. **174(2)**, 1–23.
Retrieved November 7, 2010, from: http://www.rss.org.uk/pdf/Wild_Oct._2010.pdf

Wilkerson, M. & Olson, J. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *Journal of Psychology* **131(6)**, 627–631.

Williams, A. M. (1999). Novice students' conceptual knowledge of statistical hypothesis testing. In: J. M. Truran and K.M. Truran (Eds.), Making the difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australasia (pp. 554–560). Adelaide, Australia: MERGA. ME **2000e.**03645

Zieffler, A.; Garfield, J.; delMas, R. & Reading, C. (2008). A framework to support research on informal inferential reasoning. SERJ – *Stat. Edu. Res. J.* **7(2**), 40–58. ME **2009e.**00120

APPENDIX A.
PRELIMINARY TEST BLUEPRINT

Table 3. Test Blueprint to Assess Informal Statistical Inference

(Topic Category: Informal Inference)

Category	Topics	Learning Goals	Literature
Inf-1	The concept of uncertainty	Being able to express uncertainty in making inference using probabilistic (not deterministic) language	Makar & Rubin (2009); Zieffler et al. (2008)
Inf-2	Properties of aggregates	Being able to able to reason about a collection of data from individual cases as an aggregate	Makar & Rubin (2009); Rubin, Hammerman & Konold (2006); Pfannkuch (1999)
Inf-3	Sampling variability	• Understanding • The nature and behavior of sampling variability • Understanding sample to sample variability • Taking into account sample size in association with sampling variability	Rubin, Hammerman & Konold (2006); Wild et al. (2011)

Inf-4	The concept of unusualness	Being able to understand and articulate whether or not a particular sample of data is likely given a particular expectation or claim	Makar & Rubin (2009);
			Zieffler et al. (2008);
			Liu & Thompson (2009)

Inf-5	Generalizing from a sample to a population	• Being able to predict and reason about possible characteristics of a population based on a sample of data • Being able to draw a conclusion about population from sample(s) based on the prediction	Zieffler et al. (2008)

Inf-6	Reasoning about comparison of two populations from two samples	• Being able to predict and reason about possible differences between two populations based on observed differences between two samples of data
		• Being able to draw a conclusion about comparison of two populations from two samples based on the prediction
			Wild et al. (2011);
			Makar & Rubin, (2009);
			Zieffler et al. (2008);
			Pfannkuch, (2005)

Table 4. Test Blueprint to Assess Informal Statistical Inference

(Topic Categories: Sampling distribution (SD) and Hypothesis testing (HT))

Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature
SD-1	The concepts of samples and sampling	• Understanding the definition of sampling distribution • Understanding the role of sampling distribution	A tendency to predict sample outcomes based on causal analyses instead of statistical patterns in a collection of sample outcomes	Saldanha & Thompson (2002); Saldhanha (2004); Rubin, Bruce & Tenney (1991)
SD-2	Law of Large Numbers (Sample representativeness)	Understanding that the larger the sample, the closer the distribution of the sample is expected to be to the population distribution	A tendency to assume that a sample represents the population regardless of sample size (*representativeness heuristic)*	Kahneman & Tversky; Rubin et al. (1991); Saldanha & Thompson (2002); Metz (1999); Watson & Moritz, (2000a; 2000b)
SD-3	Population distribution and frequency distributions	Understanding the relationship between frequency distribution and population distribution	Confusion between frequency distributions and sampling distributions	Sedlemeier & Gigerenzer (1997); Lipson (2003); delMas et al. (1999)
SD-4	Population distribution and sampling distributions	Understanding the relationship between sampling distribution and population distribution	Confusion between population and sampling distributions	delMas et al. (1999)
SD-5	Central Limit Theorem	• Understanding the effect of sample size in sampling distributions • Understanding how sampling error is related to making an inference about a sample mean	Lack of taking into account sample size in association with distributions of samples	Mokros & Russell (1995); Sedlemeier & Gigerenzer (1997); Tversky & Kahneman, (1974); Vanhoof et al. (2007); Schwartz, Goldman, Vye, Barron & Cognition Technology Group at Vanderbilt (1998); Wagner & Gal (1991); Well, Pollastek & Boyce (1990)

Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature
HT-1	Definition, role, and logic of hypothesis testing	Being able to describe the null hypothesis Understanding the logic of a significance test	• Failing to reject the null is equivalent to demonstrating it to be true (Lack of understanding the conditional logic of significance tests) • Lack of understanding the role of hypothesis testing as a tool for making a decision	Batanero (2000); Nickerson (2000); Haller & Krauss (2002); Liu & Thompson (2009); Vallecillos (2002); Williams (1999); Mittag & Thompson (2000)

HT-2	Definitions of *P*-value and statistical significance	Being able to recognize a correct interpretation of a *P*-value	Misconception: *P*-value is the probability that the null hypothesis is true and that (1-p) is the probability that the alternative hypothesis is true	Carver (1978);
				Falk & Greenbaum (1995);
				Nickerson (2000)

HT-3	*P*-value as a numerical probability	Understanding the smaller the *P*-value, the stronger the evidence of a difference of effect Understanding the relationship between *P*-value and standard error (Understanding that given the same mean difference, the smaller the variation in the sample statistic, the smaller the *P*-value, if all else remains the same)	Misconception: A small *P*-value means a treatment effect of large magnitude	Cohen (1994); Rosenthal (1993)

Category	Topics	Learning Goals	Misconceptions Found in Literature	Literature
HT-4	Sample size and statistical significance in HT	• Understanding larger sample sizes yield smaller *P*-values, and more statistically significant observed results, if all else remains the same	Lack of understanding the relationship between sample size and statistical significance	Wilkerson & Olson (1997)
HT-5	Evaluation of HT	Understanding that an experimental design with random assignment supports causal inference Being able to make an appropriate conclusion from a hypothesis test	Lack of interpretation of result of hypothesis testing and statistical significance	Wilkerson & Olson (1997)
HT-6	Designing a statistical test for the comparison	Being able to design a statistical test to compare two samples from a population Being able to make a conclusion from a statistical test		

APPENDIX B.
FINAL VERSION TEST BLUEPRINT

Table 5. Test Blueprint to Assess Informal Inference

(Topic Category: Informal Inference)

Category	Topics	Learning Goals
Inf-1	The concept of uncertainty	Being able to express uncertainty in making inference using probabilistic (not deterministic) language
Inf-2	Properties of aggregates	Being able to able to reason about a collection of data from individual cases as an aggregate
Inf-3	Sampling variability	• Understanding • The nature and behavior of sampling variability • Understanding sample to sample variability • Taking into account sample size in association with sampling variability
Inf-4	The concept of unusualness	Being able to understand and articulate whether or not a particular sample of data is likely given a particular expectation or claim
Inf-5	Generalizing from a sample to a population	• Being able to predict and reason about possible characteristics of a population based on a sample of data • Being able to draw a conclusion about population from sample(s) based on the prediction
Inf-6	Reasoning about comparison of two populations from two samples	• Being able to predict and reason about possible differences between two populations based on observed differences between two samples of data • Being able to draw a conclusion about comparison of two populations from two samples based on the prediction
Inf-7	Comparing two samples from two populations	• Being able to predict and reason about possible differences between two populations based on observed differences between two samples of data • Being able to draw a conclusion about two populations • Being able to take into account sample variations or sample size in relation with evidence to compare two samples

Table 6. Test Blueprint to Assess Informal Inference

(Topic Categories: Sampling distribution (SD) and Design of study (DE))

Category	Topics	Learning Goals
SD-1	The concepts of samples and sampling	• Understanding the definition of sampling distribution • Understanding the role of sampling distribution
SD-2	Sample representativeness	• Understanding importance of random sampling (recognizing biased sampling) • Law of Large Numbers (Understanding that the larger the sample, the closer the distribution of the sample • is expected to be to the population distribution)
SD-3	Population distribution, sample distributions, and sampling distribution	• Understanding the relationship between sample distribution and population distribution • Understanding the relationship between sampling distribution and population distribution
SD-4	Central Limit Theorem	• Understanding the effect of sample size in sampling distributions • Understanding how sampling error is related to making an inference about a sample mean
DE	Study design	• Understanding the logic of experimental design • Understanding difference between observational and experimental study • Understanding the purpose of random assignment in an experimental study
EV	Generalizing the results of ST Evaluation of ST	• Understanding that an experimental design with random assignment supports causal inference • Understanding that an observational design with no random assignment doesn’t support causal inference • Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, • effect size, data quality, soundness of the method, etc.

(Topic Categories: Statistical testing (ST) and Confidence interval (CI))

Category	Topics	Learning Goals
ST-1	Definitions of *P*-value and statistical significance	• Being able to recognize a correct interpretation of a P-value • Being able to calculate a numerical P-value from a given distribution of statistics • Being able to recognize a correct interpretation of statistical significance
ST-2	A statistical test for the comparison	• Being able to design a statistical test to compare two samples from two population • Designing a statistical test to compare two groups in an experiment • Being able to make a conclusion from a statistical test for comparing two groups
ST-3	Inference about a population proportion	• Designing a statistical test for the proportion given in a sample • Making a conclusion about a statistical test for the population proportion
ST-4	Inference about comparing two proportions	• Being able to set up the null model to compare two proportions • Being able to make a conclusion about a statistical test for comparing two population proportions
CI	Inference about Confidence Intervals	• Being able to interpret confidence interval in a given context • Being able to interpret the relationship between confidence interval and margin of error

APPENDIX C. EXPERTS' COMMENTS AND CHANGED IMPLEMENTED

Table 7. Experts' Comments and Changed Implemented (or Not Implemented) in the Final Blueprint

(Common suggestions)

Comments and Suggestions	Changes Made in Final Blueprint (Included or Not Included in Blueprint)
In the category of Informal inference, there is no attention to inferences about the real world or contextual knowledge. (Reviewer 1 and 2)	Added some learning goals which consider *inferential reasoning in a given context*
Formal inference (SD and ST) is too focusing on the limited population. (Reviewer 1 and 3)	Added the topics, DE (DEsign of study) and EV (evaluation of study) to get at students' understanding of characteristics of different types of study in terms of—*how to design the study* and *how to generalize the results of the study*
Need to have learning goals about understanding of effect size (Reviewer 2 and 3)	In the category EV, added the learning goal, "Being able to evaluate the results of hypothesis testing considering — sample size, practical significance, effect size, data quality, soundness of the method, etc.

(Suggestions from Reviewer 1)

Comments and Suggestions	Changes Made in Final Blueprint	
	Included or Not Included in Blueprint	Rationale for Not Included
Too focus on one type of problem, differences between groups, but almost half of the problems are about correlation problems (and regression).	Not included in the blueprint	Correlation and regression were considered as *literacy* or part of *descriptive statistics* rather than use of *inferential reasoning*
Include learning goals about "Using models in informal inferential reasoning"	In two categories, informal inference and formal inference, the learning goals about setting up the null model in a given context was added.	

Comments and Suggestions	Changes Made in Final Blueprint	
	Included or Not Included in Blueprint	Rationale for Not Included
Include using meta-cognitive awareness what inference is as opposed to performing some techniques	Not included in the blueprint	This learning goal was considered to be difficult to assess using typical test format (online format or paper-and-pencil format). Meta-cognitive awareness can be assessed through in-depth interview or individual observation.
Too focus on one type of problem, differences between groups, but almost half of the problems are about correlation problems (and regression)	Not included in the blueprint	Correlation and regression were considered as *literacy* or part of *descriptive statistics* rather than use of *inferential reasoning*
Include learning goals about "Using models in informal inferential reasoning"	In two categories, informal inference and formal inference, the learning goals about setting up the null model in a given context was added.	
Include using meta-cognitive awareness what inference is as opposed to performing some techniques	Not included in the blueprint	This learning goal was considered to be difficult to assess using typical test format (online format or paper-and-pencil format). Meta-cognitive awareness can be assessed through in-depth interview or individual observation.
Describe more explicitly about concepts like distribution, center and variation in aggregate category	In the category of *Properties of aggregates* the learning goal, *Being able to able to describe a collection of data using properties of distribution (shape, center, and variation but not necessarily using the terms)*, was added.	

(Suggestions from Reviewer 2)

Comments and Suggestions	Changes Made in Final Blueprint (Included or Not Included in Blueprint)
Need to develop a topic category on Confidence Intervals	The topic category, "Inference about Confidence Interval, CI" was added.
Need to consider data quality, soundness of the method etc.	The topic category, "Evaluation of HT (EV)", was separated out from the Hypothesis Testing categories since this topic is more about assessing how to interpret and evaluate the results from statistical testing by integrating different kinds of information in a given study (e.g., random assignment, sample size, data quality). The learning goal about, "Being able to evaluate the results of hypothesis testing (considering sample size, practical significance, effect size, data quality, soundness of the method, etc.)", was included in this EV category.
In HT-6, add designing a test to compare two groups in an experiment. You might take samples from volunteers, not from populations.	In ST-3 (changed from category of HT), the learning goal, designing a statistical test to compare two groups in an experiment, was added.
Consider including randomization and bootstrapping methods	Not included as a separate learning goals, but will be assessed in a way that items get at students reasoning of the ideas involved in randomization and bootstrap methods. Considering that hypothesis testing based on normal distribution-based approach is not the only way of statistical testing, the original category about hypothesis testing (HT) was changed to statistical testing (ST), which includes randomization or bootstrap methods.

(Suggestions from Reviewer 3)

Comments and Suggestions	Changes Made in Final Blueprint (Included or Not Included in Blueprint)
For SD-2, in addition to "how larger samples look more like the population", it is much more important "biased sampling" for sampling representativeness	The topic of "Law of Large Numbers" was changed to "sample representativeness" to assess whether students realize the importance of unbiased sampling (quality of samples) in addition to a large number of a sample (quantity of samples)