# DEVELOPMENT AND VALIDATION OF THE STATISTICS TEACHING INVENTORY (STI)

Jiyoon Park
University of Minneapolis, United States of America
parkx666@umn.edu

*The reform movement in science, technology, engineering, and mathematics has prompted calls for research to provide a better understanding of factors related to desired learning outcomes in introductory statistics courses. However, little is known about instructor factors involved in teaching statistics. This study describes the development and validation of the Statistics Teaching Inventory (STI) designed to assess the practices and beliefs of teachers of introductory statistics courses across the disciplines. Survey data were collected from 101 instructors from different institutions in the United States. Reliability coefficients were high (>0.75) for two subscales (practice and beliefs) and item-total correlations were generally at an acceptable level (>0.30). Summaries of individual response patterns and the relationship between the beliefs and practice subscales are presented. The study concludes with limitations related to the sample size, sample characteristics, and psychometric properties based on summated scores, and suggests future research.*

## BACKGROUND

With increasing calls for reform in undergraduate education in science, technology, engineering, and mathematics (STEM) disciplines, there has been a corresponding increase in attention to teaching statistics at the introductory level. To improve the teaching and learning statistics, statisticians and statistics instructors concerned with teaching statistics reexamined the introductory course and offered recommendations for revision (Cobb, 1992; Garfield, 1995; Moore, 1997; Velleman & Moore 1996). Many projects involving the reform movement in statistics education commonly addressed the need for changes in pedagogy, content, technology, and assessment of introductory statistics courses (See, for example, Moore, 1997).

Integrating the ideas of what and how to teach statistics, the recommendations for statistics education in K-12 levels and introductory college level were documented in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) report. It describes the student learning outcomes that should be obtained from statistics courses in terms of statistical literacy, statistical reasoning, and statistical thinking. This report also integrates the ideas of the report by Cobb (1992) that provides ways of teaching statistics for introductory statistics instructors. The guidelines for the teachers to be more effective in helping students achieve the learning outcomes are summarized with six recommendations (See http://www.amstat.org/education/GAISE/GAISECollege.htm for the detailed descriptions).

In 2005, The GAISE report was approved by the ASA as a guideline for teaching introductory statistics. Despite the reform efforts, however, research shows that students still lack understanding of statistical concepts (See, for example, delMas et al., 2007). Moreover, a survey for statistics instructors revealed that teachers have resistance to recommended ways of teaching (see, for example, Garfield & Ben-Zvi, 2008) despite the positive effects of the reformed teaching approach (See, for example, Chick & Watson, 2002). From this current status of teaching and learning statistics, this study was conducted to find the link between student learning outcomes and teacher's approach to teaching.

The project funded by a grant from NSF developed an instrument, the Statistics Teaching Inventory (STI), designed to assess practice and beliefs of teachers in introductory statistics courses. This project was a part of a larger project that will develop and pilot an instrument and integrate it into a database to explore the relationship between teaching and student learning in introductory statistics courses. As an initial step, in the current study, we attempt to validate the appropriateness of the STI by examining the psychometric properties as well as the summary statistics of the pilot data. The validation study of the instrument will help design the next version of the STI to better capture teaching approaches in statistics. The analyses of the pilot data will also provide the interpretations of the standing scores for the samples in each measure of the instrument.

In C. Reading (Ed.), Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

## INSTRUMENT AND METHODS

To measure the degree to which instructors' teaching approaches are reformed the GAISE recommendations were used as a framework in the development of the STI. The first version of the STI with 102 items was revised and pared down to a smaller set of items based on feedback from members of the statistics education community such as the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) and the Research Advisory Board (RAB) of CAUSE. Focus groups were conducted with two different types of faculty to obtain feedback on the smaller set of STI instrument: one from statisticians teaching in other disciplines outside statistics and mathematics such as department of psychology and school of business, and the other from education researchers in other postsecondary STEM disciplines. After the focus group with faculty from the two different disciplines, online pilot testing was followed by focus group interviews of statistics educators in the greater Twin Cities area in Minnesota. The resulting version of the STI was administered to 101 participants of the 2009 US Conference on Teaching Statistics (USCOTS) to gather data to examine reliability and scale construction. To obtain the validity evidence of the STI, face-to-face and phone interviews were performed. Course syllabi and course materials were also collected from each interviewee and analyzed as the evidence for the STI scale scores.

In the current study, the data from 101 instructors from different institutions and departments who participated in the STI were analyzed. The 101 instructors were from university level (49.5%), 4-year college level (37.6%), and 2-year college level (12.9%). They were teaching statistics in Mathematics or Statistics departments (73.3%), Liberal Arts, Education, Psychology, or Sociology related departments (13.9%), and Business, Engineering, Biology, or science related departments (13%).

## MEASURES AND CODING

The STI is a scale consisting of 50 multiple-choice items developed by delMas, Garfield and Zieffler (2009). Four parts of the STI, *Teaching Practice* (Part 1), *Assessment Practice* (Part 3), *Teaching Beliefs* (Part 4), *and Assessment Beliefs* (Part 5), assess how much the instructors' beliefs and practice are reformed (or traditional) in teaching introductory level of statistics in a perspective of GAISE. Part 2 of the STI involves questions about general characteristics of the course such as the number of students in the class or the mathematical prerequisite. Part 6 includes items about demographic information such as classifications of the institution and the department.

The original data were recoded to reflect the constructs to be measured in each part as well as theoretical and empirical aspects of measurement. Issues regarding how to make reasonable scale scores were discussed during several meetings with the instrument developers as well as consultation with a professional in measurement. The responses for traditional items in the four main parts were reversely coded so that a higher score for an item indicates a more reformed approach. Since the instrument employs the mixed-item format including different response scales (dichotomous, 4-point, and 5-point scale), each response of the questionnaire was recoded to the scale with a partial score between 0 and 1. Use of the partial scores of a 0 to 1 scale across the items and scales consistently allowed the scores to be compared, item by item and part by part. For the *Beliefs* part, the "Undecided" response option was recoded as a "missing value" because we considered "Undecided" to be on a different scale than the other four ordinal responses ("Strongly Disagree" to "Strongly Agree") which indicate the degree of respondents' agreement to each item. We also considered that the respondents who chose the "Undecided" option might have diverse reasons for the choice, which could not be identified from the item analysis.

## RESULTS

The STI data analyses consist of two aspects–item analysis and scale analysis. The measurement validation was based on the Classical Test Theory (CTT) which provides the reliability coefficient, individual item properties, and scale scores. In the *item analysis*, mean score and discrimination of each item were analyzed. The corrected item-total correlation was used as a measure of discrimination. In the *scale analysis*, internal consistency and mean scores for the entire questionnaire as well as two subscales (practice and beliefs) were analyzed. As a measure of internal consistency, Cronbach's α was used to evaluate how consistently the STI measures a teacher's approach to teaching. The mean scores were used to characterize the instructors' overall teaching approaches in practice and beliefs.

*Item analysis*

Mean scores for most of the items in the *Practice* part were greater than 0.5 except for the items describing "use of teacher presentation", "group discussion" and "homework type problems". The standard deviations for the mean scores ranged from 0.19 to 0.50. The item-total correlations for the *Practice* items were greater than 0.30, except for the four items about "homework-type problems", "using a variety of types of assessment", "statistical literacy", and "technology". Most of the *Beliefs* items also had mean scores over 0.5 except for the five items which describe "probability rules", "theoretical probability distributions", "statistical tables", and "traditional assessment". The standard deviations for the mean scores ranged from 0.15 to 0.32 which is generally smaller than the standard deviations of the *Practice* items. The item-total correlation for each item in the *Beliefs* part was greater than 0.30 except for three items: "fewer topics in greater depth", "statistical literacy", and "purpose of student assessment". The item examples with the highest and lowest mean score in each part and the other properties of the item are presented in Table 1.

Table 1. Item examples with the highest and the lowest mean score, and the item properties

| Subscale | Item | N | Mean (SD) | Corrected Item-Total Correlation | Standardi-zed Alpha if Item Deleted |
|---|---|---|---|---|---|
| Teaching Practice | The need to base decisions on evidence. | 101 | 0.77 (0.22) | 0.36 | 0.76 |
| | Teacher presentations are used to help students learn statistics.[a] | 101 | 0.22 (0.19) | 0.40 | 0.75 |
| Assessment Practice | My assessments evaluate students' ability to interpret results of data analysis. | 101 | 1.00 (0.00) | NA | NA |
| | My assessments evaluate students' abilities to use formulas to produce numerical summaries of a data set.[a] | 101 | 0.48 (0.50) | 0.62 | 0.74 |
| Teaching Beliefs | Technology tools should be used to illustrate most abstract statistical concepts. | 90[b] | 0.84 (0.19) | 0.47 | 0.79 |
| | Rules of probability should be included in an introductory statistics course.[a] | 83[b] | 0.46 (0.30) | 0.56 | 0.79 |
| Assessment Beliefs | It is important to assess student statistical literacy. | 100[b] | 0.89 (0.15) | 0.21 | 0.81 |
| | Traditional assessments should be used to evaluate student learning.[a] | 98[b] | 0.24 (0.18) | 0.30 | 0.80 |

a. Traditional items: items were reverse-coded: higher values close to 1 indicate reformed approach and lower values close to 0 indicate less-reformed approach in each subscale.

b. Total N = 101. The different sample size for each part resulted from coding the "Undecided" response option as a "missing value".

*Scale analysis*

The mean scores for the separate scales of *Practice* and *Beliefs* as well as the combined scale are higher than the neutral value of 0.5. It is important to note that all of the respondents in our study participated in the USCOTS which is a conference that aims to facilitate instructors incorporating ideas of teaching into existing courses and programs. Since the mixed item format included a different number of response options for items in the same scale, standardized Cronbach-α was provided as a measure of internal consistency (Santos, 1999). The overall Cronbach-α was high (0.87), which is sufficient for individual-level measurement and well above the suggested cutoff of 0.70 for group-level measurement. The Cronbach-α's for the *Practice* items and the *Beliefs* items were 0.76 and 0.80, respectively.

Table 2. Practice and beliefs scales

| Subscale | Mean (SD), N=101 | | | Cronbach α |
|---|---|---|---|---|
| | Teaching | Assessment | Total | |
| Practice | 0.58<br>(0.12) | 0.74<br>(0.17) | 0.66 (0.13) | 0.76 |
| Beliefs | 0.64<br>(0.15) | 0.62<br>(0.12) | 0.63 (0.12) | 0.80 |
| Total | | 0.63 (0.12) | | 0.87 |

## CONCLUSIONS AND DISCUSSIONS

The psychometric validation of the STI was supported through the high values of Cronbach-α for the entire scale (0.87), and for two subscales of *Practice* and *Beliefs* (0.76 and 0.80, respectively). These results support the utility of the STI in measuring instructors' approaches to teaching introductory statistics. It appears that the instructors sampled in this study use a moderately reformed approach in teaching statistics with mean scores of 0.66 and 0.63 for each part. The difference between the mean scores for the *Teaching Practice* items (0.58) and the *Assessment Practice* items (0.74) indicates that the instructors tend to use more reformed approaches in assessment than in teaching. This may indicate that these instructors find it more feasible to implement reformed approaches in assessment than in actual teaching. The instructors showed almost the same degree of implementing reform approaches in teaching and assessment (0.64 and 0.62, respectively) for the *Beliefs* part.

Despite the validation evidence shown above, this study has some limitations. First, a larger sample is needed to make statistical and psychometric analyses more valid. Secondly, the subjects sampled in this study are biased in that they were recruited through a conference sponsored by the statistics education research field that seeks to improve teaching statistics. The fact that 71% of the subjects were aware of the GAISE recommendations indicates that many instructors surveyed might have been already familiar with what GAISE aims for in teaching statistics. There are also limitations regarding the coding of items and the scaling of scores in order to deal with the mixed-item-format.

## REFERENCES

Chick, H., & Watson, J. (2002). Collaborative influences on emergent statistical thinking: A case study. *Journal of Mathematical Behavior, 21*, 371-400.

Cobb, G. (1992). *Teaching statistics*. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (MAA Notes No. 22), 3-43.

Cobb, G. (1993). Reconsidering statistics education: A national science foundation conference. *Journal of Statistics Education*, *1*(1).

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Educational Research Journal, 6*(2), 28–58

delMas, R., Garfield, J., & Zieffler, A. (2009). The National Statistics Teachers' Practice and Beliefs Project. *Talk presented at the 2009 Joint Statistical Meetings*, Washington, D.C.

Garfield, J. (1995). How students learn. *International Statistical Review, 63*, 25-34.

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students Statistical Reasoning: Connecting Research and Teaching Practice*. Dordrecht, The Netherlands: Springer.

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123-165.

Santos, R. (1999). Cronbach's alpha: A tool for assessing the reliability of scales, *Journal of Extension, 37*(2), Retrieved January 5, 2010, from http://www.joe.org/joe/1999april/tt3.php.

Velleman, P., & Moore, D. (1996). Multimedia for teaching statistics: Promises and pitfalls. *The American Statisticians*, *50*, 217-225.