International Journal on Social and Education Sciences
Volume 1, Issue 2, 2019 ISSN: 2688-7061 (Online)

Attitudes toward Using and Teaching Confidence Intervals: A Latent Profile Analysis on Elementary Statistics Instructors

Hyung Won Kim
University of Texas Rio Grande Valley, USA, hyung.kim@utrgv.edu

Woo Jin Kim
University of Texas Rio Grande Valley, USA

Aaron T. Wilson
University of Texas Rio Grande Valley, USA

Ho Kyoung Ko
Ajou University, South Korea

Abstract: The use of confidence intervals (CIs) for making a statistical inference is gaining popularity in research communities. To evaluate college statistics instructors' readiness to teach CIs, this study explores their attitudes toward teaching CIs in elementary statistics courses, and toward using CIs in inferential statistics. Data were collected with a survey that classifies instructors' attitudes on the basis of three previously established pedagogical components: affective, cognitive, and behavioral. Based on the survey responses from 270 participants, we created three profiles (subgroups) via latent profile analysis, and identified each profile's pattern of attitudes toward CIs and common characteristics of the instructors that fit each profile. In addition, we compared the profiles across groupings created by six variables: gender, academic background, statistics teaching experience, subject preference, degree level, and desire to improve teaching. The results of the latent profile analysis support three profiles within the population of statistics instructors, and the results of the comparative analysis of teacher characteristics indicate that the six variables are moderate to strong predictors of the grouping of the sample into three profiles.

Keywords: Confidence intervals, Inferential statistics, College instructors of statistics, Latent profile analysis, and Non-cognitive factors

Introduction

In statistics, the process of null hypothesis significance testing (NHST) uses *p*-values to make statistical inferences regarding research results. This process has been the overwhelming choice for quantitative data analysis in the social sciences for many years (Fidler & Loftus, 2009). However, an increasing number of scholars in statistics education and in quantitative research methods in psychology consider the *p*-value an inadequate measure of evidence for hypothesis testing in inferential statistics at the elementary statistics level (Hubbard & Lindsay, 2008; Lindley, 1999; Marden, 2000; Nelder, 1999). Recently, many statistics reformers have studied the adequacy of confidence intervals (CIs) in inferential statistics (e.g., Beyth-Marom, Fidler, & Cumming, 2008; Cumming & Finch, 2005; Kline, 2004).

In the pedagogical context, Schmidt and Hunter (1997), for example, claimed that students might have greater success learning how to use CIs than learning how to conduct NHST, and, further, that CIs enable researchers to make better inferential decisions than the *p*-values used in NHST. In recognition of the advantages of using CIs in inferential statistics, reformers of statistics practices have called for wider usage of CIs in preference to *p*-values (Cumming & Finch, 2001). As a result, interval estimates are gaining more attention and becoming more popular in the quantitative research community (Cumming & Finch, 2005), leading to a need for more studies on the usage and teaching of CIs.

The literature on this topic, although still scant, indicates that a number of statistics educators are actively exploring cognitive factors in students' attainment of knowledge of CIs, particularly misconceptions of CIs commonly developed by statistics students (e.g., Fidler, 2005; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Kalinowski, 2010). The role of teachers' noncognitive factors in improving the cognitive performance of

students has also been investigated in the statistics education community (e.g., Gal & Ginsburg, 1994; Kim, Wang, Lee, & Castillo, 2017; Martins, Nascimento, & Estrada, 2012). A recent study by Kim, Wilson, and Ko (2018) is the first to address three different aspects—affective, cognitive, and behavioral—of the attitudes of college instructors of elementary statistics toward teaching CIs in elementary statistics and toward using CIs for inferential statistics. Expanding on Kim et al.'s (2018) work, the current study aims to classify the population of statistics instructors into groups, which we call *profiles*, to identify each profile's pattern of attitudes toward CIs, and to identify common characteristics of the instructors that fit each profile.

Research Background

In this section, we describe the known advantages of learning and using CIs in inferential statistics, discuss the pedagogical components of attitudes and instructor characteristics, and present the study's research questions.

Advantages of Learning and Using CIs in Inferential Statistics

The *p*-value methods have been widely accepted and used as a primary tool for data analysis in quantitative studies that report their findings via inferential statistics in many areas of social science and other research fields. Accordingly, null hypothesis significance tests (NHSTs), of which the *p*-value is an essential component, are heavily emphasized in the discussion of inferential statistics in college level elementary statistics courses, and the *p*-value has been regarded as a culminating topic in these courses. Recent studies, however, indicate declining favor and use of *p*-values in research communities, and some statistics reformers emphasize the advantages of using CIs for research purposes, and the need to teach CIs as a tool for inferential statistics in elementary statistics classes. In the following subsections, we describe three known advantages of learning and using CIs over *p*-values: they are easier to access, provide more information, and promote meta-analysis.

Easier Access

Students are known to develop misconceptions in learning the concept of *p*-values. Castro Sotos, Vanhoof, Van den Noortgate, and Onghena (2009) documented the severity of misconceptions, and misinterpretations associated with the learning of *p*-values and hypothesis tests. In contrast, according to Cumming and Finch (2001), the point and interval information provided in CIs supports substantive understanding and interpretation. Empirical studies conducted by Fidler and Loftus (2009) support this claim, concluding that visually presented CIs ―result in fewer misconceptions than p-values used with NHST" (p. 32). In particular, CIs help make sense of statistical non-significance (Cumming & Finch, 2001, 2005; Kalinowski, 2010). The empirical studies of Fidler and Loftus indicate that ―CIssubstantially alleviate the misinterpretation of statistically nonsignificant results as evidence for the null hypothesis" (p. 36).

More Information

While remaining within a familiar frequentist framework, where inferential conclusions are drawn from the frequencies present in the data, CIs offer more information than *p*-values. According to Cumming and Finch (2001), there is a direct link between CIs and null hypothesis significance testing (NHST):

> Noting an interval excludes a value is equivalent to rejecting a hypothesis that asserts that value as true—at a significance level related to *C*. A CI may be regarded as the set of hypothetical population values consistent, in this sense, with the data. (p. 534)

CIs give additional information. According to Cumming and Finch, the precision information provided by CIs can be estimated before conducting an experiment, and the width may be used to guide the choice of design and sample size. Moreover, after an experiment is completed, CIs give information about precision that may be more useful and accessible than a statistical power value (Cumming & Finch, 2001). Although statistical power and effect sizes help boost the meaning of *p*-values and their interpretation, graphical representations of CIs provide a visual assessment of data that a reader can use to understand underlying patterns.

Promoting Meta-analysis

CIs are helpful when combining evidence from multiple experiments because they promote meta-analytic thinking. According to Fidler and Loftus (2009), CIs can be useful for investigating multiple hypotheses on a single scale because they do not presume any single null hypothesis (p. 29). Therefore, CIs emphasize the idea of uncertainty, and they more readily allow readers to assess data at varying degrees than do the *p*-values used in NHST, which lead strongly toward a dichotomous conclusion of either rejecting or failing to reject the null hypothesis. In this sense, CIs can lead readers to look at data in a research paper more holistically, promoting meta-analytical thinking. According to Cumming and Finch (2001):

> CIs are useful in the cumulation of evidence over experiments: They support meta-analysis and meta-analytic thinking focused on estimation. This feature of CIs has been little explored or exploited in the social sciences but is in our view crucial and deserving of much thought and development. (p. 534)

The authors of this paper believe that it is important for instructors to be aware of these advantages of CIs, to be ready to use CIs as a tool for inferential statistics, and to view them as a valued topic to emphasize in elementary statistics courses.

Pedagogical Components of Attitudes and Instructor Characteristics

A taxonomical construct that classifies attitude into three pedagogical components—affective, cognitive, and behavioral—has been widely accepted in studies of noncognitive factors in psychology and education (e.g., Aiken, 1970; Gómez-Chacón, 2000; Olson & Zanna, 1993). Recently, this taxonomical construct has been used to address pre- and in-service teachers' attitudes toward statistics (e.g., for preservice elementary school teachers, Estrada, 2002 and Martins et al., 2012; for college instructors, Kim et al., 2017). According to Estrada (2002) and Martins et al. (2012), the three components can be defined as follows:

- Affective component: one's feelings about the objects in question;
- Cognitive component: one's self-perception or beliefs about the objects in question; and
- Behavioral component: one's inclinations to act in a particular way about the objects in question

This study also draws upon this taxonomical construct. However, rather than using the three pedagogical components to distinguish different types of attitudes, we use them to cover widely used pedagogical domains of attitudes toward teaching and using CIs.

Using the same taxonomical construct and the same definition of each component, Kim et al. (2018) developed a survey to explore each of the three aspects (affective, cognitive, and behavioral) of instructors' attitudes toward CIs, and to learn how the attitudes are explained by three factors: *gender*, *academic background*, and *statistics teaching experiences*. Their findings indicate that *gender* is a moderate predictor while *academic background* and *statistics teaching experience* are strong predictors of elementary statistics instructors' (ESIs) attitudes toward CIs. Specifically, they showed that the ESIs with a statistics background at the graduate level held significantly more positive attitudes toward CIs than those without such a background, and the ESIs with four or more years of teaching experience had more positive attitudes toward CIs than their counterparts with less than four years of experience. These results suggest the likelihood of distinct subgroups within the population of ESIs.

Extending the study of Kim et al. (2018), the study in this paper intends to identify existing subgroups within the ESI population, and to explain the characteristics of ESIs in each subgroup using the same three potential factors (gender, academic background, statistic teaching background) along with three additional factors: degree level, subject preference, and self-perception of the need for professional development. The specific questions we address are:

- In the target population, how heterogeneous are the participants in terms of attitudes toward confidence intervals?
  - How many distinct groups exist in the target population?
  - What attitudinal patterns do ESIs exhibit toward using CIs and teaching CIs?
- What relationships exist between the attitudinal pattern groups (i.e., profiles) and the six variables of gender, academic background, statistics teaching experience, degree level, subject preference, and desire to improve teaching?

Methods

To measure ESIs' attitudes toward CIs, we used a survey that was designed for this purpose. See Kim et al. (2018) for brief results related to taxonomical constructs and teaching contexts.

Description of Survey and Data Collection

The survey has three parts. Part I collected biodata of the participants, and provides the information on the six teacher characteristic variables we explore in this study: gender, academic background, statistics teaching experience, subject preference, degree level, and desire to improve teaching. Part II contains 12 items that address ESIs' attitudes toward using and teaching CIs without making explicit connections with the aforementioned advantages of CIs. Part III contains 18 items, which address ESIs' attitudes toward CIs in relation to such advantages. This study uses the participants' responses to the items in Parts I and II only.

The questionnaire of Survey Part II draws on the taxonomical construct of the three pedagogical components of attitude (Table 1, columns [A], [C], and [B]). In addition, the questionnaire items are framed so as to assess (1) two views—one of attitudes toward CIs in place of *p*-values and the other of attitudes toward CIs regardless of *p*-values (Table 1, rows [a] and [b]), and (2) two contexts—the general context and the teaching context (the sub-columns under columns [A], [C], and [B] in the table).

Table 1. Survey Items Framed by the Taxonomical Constructs

	Affective (A)		Cognitive (C)		Behavioral (B)	
	General	Teaching	General	Teaching	General	Teaching
(a) In comparison with *p*-values (COM)	I8	I9	I12	I7	I11	I10
(b) Independent of *p*-values (IND)	I1	I3	I5	I4	I6	I2

The 12 items in Part II have a Likert-scale format. To increase the effect of responses, it uses a 7-point scale instead of the usual 5-point scale (from 1 "strongly disagree" to 7 "strongly agree"). To avoid apparent acquiescence, the survey has a mixture of positive and negative items; seven items are phrased positively and five negatively. The response scores of the negative items were reversed for the analysis. The 12 items appear in the appendix.

The data were collected via an online survey system during the summer of 2017. An invitation to take the survey was sent to 10,096 faculty in mathematics or statistics departments at two- and four-year colleges in 24 states across the United States (i.e., all faculty members in the relevant departments at all two- or four-year colleges we could identify in the 24 states). While 292 responded, the analysis was conducted using the responses of 270 participants who answered more than 80% (10 items) of the 12 items in Part II of the survey. The distributions of the participants based on each of the six variables—gender, academic background, statistics teaching experience, subject preference, degree level, and desire to improve teaching—are shown in the Results section.

Validity

The survey was designed to ensure three types of validity: *content*, *face*, and *construct validity*.

Content Validity

Content validity refers to the extent to which the items on a measure represent all facets of the targeted construct (Haynes, Richard, & Kubany, 1995) or the extent to which an instrument adequately samples the research domain of interest (Wynd, Schmidt, & Schaefer, 2003). To establish content validity, a group of four math and statistics educators (including three of the four authors of this paper) had back-and-forth discussions during the survey development phase. The discussion focused on whether the 12 items in Part II of the survey covered: (1) the three domains determined by affective, cognitive, and behavioral aspects, (2) the two views—of attitudes toward CIs in place of *p*-values and of attitudes toward CIs regardless of *p*-values—and (3) the two contexts—the general context and the teaching context.

Face Validity

Face validity is generally defined as the degree to which a test (or assessment) seems to measure what it is supposed to measure from a participant's perspective. That is, it refers to the relevance of a test as it appears to test participants. To secure face validity, the research team consulted with two instructors of elementary statistics; one holds a PhD in a statistical science, and had taught various statistics courses for over three years at the time of the consultation; and the other holds an MS degree in statistics, and had taught elementary statistics for eight years at the time. They both gave advice on technical uses of statistics terms to ensure that the terms were used appropriately in the survey items and that the survey items were phrased such that they ask what the research team intended to ask. Their suggestions were noted and discussed in this paper.

Construct Validity

Construct validity is the extent to which a test measures a theoretical construct or trait (Anastasi & Urbina, 1997). As described above the survey used in this study was based on a taxonomy of three components that together comprise pedagogical attitudes towards CIs. Therefore, for construct validity, we did not treat these three components separately in this analysis, but consider the internal consistency of the 12 items together. Based on the responses, the instrument demonstrated acceptable internal consistency, with a Cronbach's alpha of 0.751. Factor analysis was not considered because this study's analysis considers the 12 items in the survey either holistically or individually, but not as distinct sets.

Results

This study's intention is to create profiles (or define subgroups) of the given group of ESIs using latent profile analysis, to explore each profile's attitudinal patterns in regard to teaching and using CIs, and to characterize the ESIs of each profile in terms of the six aforementioned teacher characteristics. All significance tests were conducted at the commonly used significance levels of α = .05, .01, and .001. Due to the limitations involved in making dichotomous conclusions based on *p*-values (Gelman, 2013), we also considered α = .10 at times (for example, when we tested possibilities for potential factors that might characterize the profiles). Furthermore, we provide 95% two-sided confidence intervals to describe the distribution of the scores of each item for the identified profiles.

Subgroups of the Target Population

The results of the LPA are shown in Table 2. The five indices (or criteria) under the "Fit statistics" column in Table 2 measure the fit of different numbers of profiles (1 through 4 in the "Model" column in Table 2) to the given sample. The measures under each index show the degree to which the associated index supports each number of profiles. For example, the second number (11103.304) in the AIC column measures how well AIC agrees with dividing the given sample into two profiles. AIC, BIC, and SABIC support the number of profiles associated with the lowest measure of that index, which shows which model fits well using as few parameters as possible (Tein, Coxe, & Cham, 2013).

Table 2. Number of Profiles Suggested by the Five LPA Criteria

Model	Fit statistics					Profile membership distribution			
	AIC	BIC	SABIC	BLRT	Entropy	1	2	3	4
1 profile	11613.038	11699.401	11623.304	-	-	270			
2 profiles	11103.304	11236.446	11119.131	-5782.519***	.945	54	216		
3 profiles	10960.805	**11140.726**	10982.192	-5514.652***	**.962**	**49**	**209**	**12**	
4 profiles	**10859.661**	11806.362	**10886.609**	**-5430.402*****	.837	45	109	12	104

Notes. (1) AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; and BLRT = bootstrapped likelihood ratio test, and (2) $^{*}\ p < .05$; $^{**}\ p < .01$; and $^{***}\ p < .001$

BLRT determines an optimal number of profiles by comparing a model, via statistical inference, that supports G profiles against a model that supports $G-1$ profiles for each $G \geq 2$. Thus, the three BLRT measures -5782.519^{***}, -5514.652^{***}, and -5430.402^{***} indicate, respectively, that a two-profile model provides a significantly higher degree of fit than a one-profile model, a three-profile model provides a significantly higher

degree of fit than a two-profile model, and a four-profile model provides a significantly higher degree of fit than a three-profile model, all at the significance level of $\alpha = .001$. Therefore, BLRT most strongly supports a four-profile model. For entropy, the higher the measure is, the more strongly it supports the associated number of profiles.

The measures in Table 2 indicate that while AIC, SABIC, and BLRT support a model of four profiles within the population represented by the given sample, both BIC and entropy support a model of three profiles. In this study, we regard three profiles as more optimal than four profiles for further analysis for several reasons: (1) BIC, which is in general more widely used than the other four indices (Tein, Coxe, & Cham, 2013), supports three profiles; (2) four profiles are least supported by both BIC and entropy; and (3) when four profiles were assumed, we found significant overlaps on many of the item scores between Profiles 2 and 4. The last four columns of Table 2 (under "Profile membership distribution") suggest the size of each profile for each assumed number of profiles. If three profiles are assumed within the sample, the sample sizes would be: n_1 = 49, n_2 = 209, and n_3 = 12 (shown in the shaded part of the table).

Under the three-profile assumption, we further tested the latent class patterns based on estimated posterior probabilities (Table 3). For example, in Table 3, the three numbers in columns 1, 2, and 3 in the Profile 1 row mean that when an individual belongs to Profile 1, the probability that the latent test identifies that individual as belonging to Profiles 1, 2, and 3 is, respectively, .984, .016, and .000. In general, the profiling is considered *acceptable* if an individual's probability of belonging to his or her own profile is greater than .7 for all profiles. The measures of .984, .985, and .967 in Table 3 indicate the probabilities that individuals in Profiles 1, 2, and 3 are classified in Profiles 1, 2, and 3, respectively.

Table 3. Test of the Latent Class Patterns Based on Estimated Posterior Probabilities

Profiles	*n*	1	2	3	Sum
Profile 1 (P1)	49	**.984**	.016	.000	1
Profile 2 (P2)	209	.015	**.985**	.000	1
Profile 3 (P3)	12	.010	.023	**.967**	1

Profile Characteristics

The study further considers the extent to which the three profiles are different from one another for each of the 12 items in Part II of the survey. While the profiles were determined by the differences shown in the individuals' responses to the 12 items, the degree of difference on each item varied. The mean differences are summarized graphically in Figure 1. For convenience, we refer to Profiles 1, 2, and 3, respectively, as P1, P2, and P3. Note that P2 contains the majority of the sample (n_2= 209), and that P2 scored higher than P1 or P3 for most items. The only two items where P2 scored lower than some other profile are Items 9 and 11, for which the difference seems nonsignificant. For the other 10 items, the mean scores of the individuals in P2 seem clearly higher than at least one of the other two profiles' mean scores. In this section, we report the group differences for each of the 12 items.

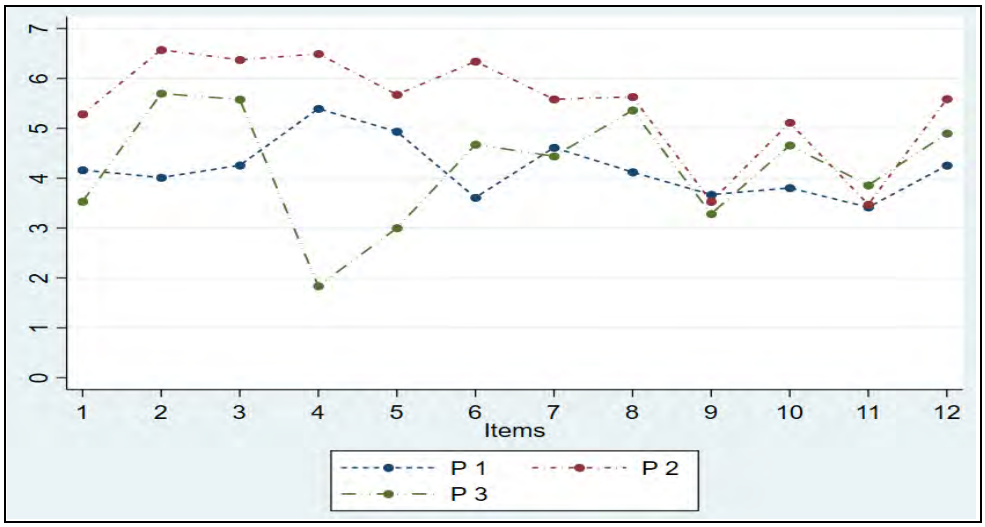

Figure 1. Mean Profile Scores in the Three Latent Profiles

Figure 1 shows the profiles' mean differences on each item by point estimate. Next, 95% interval estimates, and the profile differences based on the interval estimates, are shown in Figure 2. The graphs in Figure 2 show that Items 1, 2, 4, 5, and 6 contain a pair of intervals with gaps. This indicates the possibility of profile differences for those items. In particular, Item 4 contains a wide gap between P3 and the other profiles. Also, based on the point estimates and the interval lengths of the scores for the three profiles, we describe P1, P2, and P3, respectively, as the ~~"s~~table-score group," the ~~"h~~igh-score group," and the ~~"d~~iverse-score group." The scores of P1, the stable-score group, show more consistency over the 12 items compared to the other two profiles (Figures 1 and 2); the scores of P2, the high-score group, are higher than the scores of the other two profiles for most items (Figure 1); and the scores of P3, the diverse-score group, fluctuate across the items more than those of the other two profiles (Figures 1 and 2).

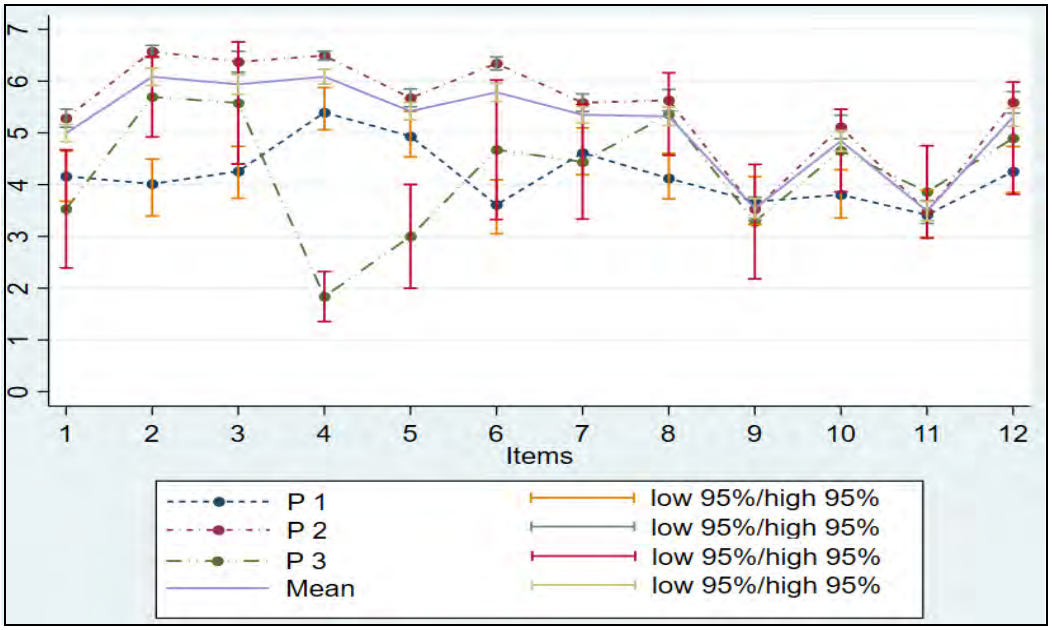

Figure 2. Plot of the Latent Profile Analysis with 95% Confidence Intervals

While the 12 items served as variables that determine the profiling, they also serve to compare different attitudinal patterns of the profiles, and to describe the profile differences. Table 4 shows the results of the profile difference tests that used an *F*-test statistic (Column B), post-hoc analysis (Column C), and effect sizes (Columns D and E) for each of the 12 items.

Table 4. Profile Differences on Each Item

(A)	(B)	(C)	(D)	(E)
Items	*F*-value	post-hoc analysis (Scheffé)	η^2	ω^2
I1	25.504***	1 < 2, 2 > 3	.161	.154
I2	147.797***	1 < 2, 2 > 3, 3 > 1	.525	.521
I3	49.842***	1 < 2, 3 > 1	.273	.267
I4	277.310***	1 < 2, 2 > 3, 3 < 1	.677	.673
I5	29.747***	1 < 2, 2 > 3, 1 > 3	.182	.176
I6	162.839***	1 < 2, 2 > 3, 1 < 3	.549	.545
I7	15.929***	1 < 2, 2 > 3	.107	.100
I8	24.819***	1 < 2, 3 > 1	.157	.150
I9	.185	-	.001	.000
I10	13.509***	1 < 2	.092	.085
I11	.497	-	.004	.000
I12	18.388***	1 < 2	.121	.114

* $p < .05$; ** $p < .01$; *** $p < .001$

In Table 4, Column (A) lists the 12 items. Column (B) shows the results of a one-way ANOVA test that we conducted to learn group differences using an *F*-test for each item. The results are shown in Column (B). For each of the 10 items that show statistical significance (all but Items 9 and 11), we further tested where the differences lie among the three profiles via post-hoc analysis using Scheffé's method without assuming equal

variance. Column (C) shows the outcomes of one-tailed tests. For example, "1 < 2" for Item 1 indicates that the mean score of Profile 2 is significantly higher than that of Profile 1 at the significance level of $\alpha = .001$ for the item. For the effect sizes, we report eta squared (η^2: Column D) and omega squared (ω^2: Column E). The effect sizes draw attention to the items that displayed greater group differences. The effect sizes of the group differences shown in I2, I3, I4, and I6 were more significant than those for the other items with significant differences. We include I3 with reservation, due to the significantly lower effect sizes of the item ($\eta^2 = .273$ and $\omega^2 = .267$) compared to the other three. While *F*-value alone indicates that group differences are statistically significant for 10 items (all but I9 and I11) at $\alpha = .001$, the *F*-values combined with the effect sizes indicate that the group differences are more significant in these four (I2, I3, I4, and I6). The outcomes from point estimation, interval estimation, and *F*-tests along with the effect sizes *not only* indicate significant group differences on the four items of I2, I3, I4, and I6 *but also* suggest the slimness of the possibility of such significance on the two items of I9 and I11. To illustrate the characteristics of each profile, we draw upon these six items, which are listed in Table 5. In the table, we refer to the four items (I2, I3, I4, and I6) with greater difference effect sizes between profiles as *significant items*, and the two items (I9 and I11) smaller effect sizes as *nonsignificant items*.

Table 5. Four Items with Significant Differences and Two Items Without

Impact Status	Items	Statements
Four significant items	I2	I rarely explain CIs in my classes as a statistical tool for making decisions about data.
	I3[a]	I feel uncomfortable teaching the underlying concept of CIs.
	I4	CIs should be discussed as a method for making statistical inferences in an introductory statistics course curriculum.
	I6	I rarely approach inferential statistics using CIs.
Two nonsignificant items	I9	I feel more comfortable teaching CIs than p-values.
	I11	I tend to treat statistical inference problems in terms of CIs.

[a] The differences on Item 3 had a much lower effect size than the other three items.

Regarding the four significant items, note that all four of these items are from the independent view, indicating attitudes toward CIs without comparison to *p*-values. We further note that three of them (I3, I4 and I2) are from the teaching context, rather than the general context. Regarding the two nonsignificant items (I9 and I11), the graphs in Figures 1 and 2 as well as the *F*-values (.185 for I9 and .497 for I11) show that the group difference for each of the two items is not significant at the level of α = .05. This evidence is further supported by the ranking given by Cramer's V (i.e., effect size; Cramer, 1946). Note that both these items are from the in-comparison view as opposed to the independent view, framing CIs in comparison to *p*-values.

Characteristics of the Grouping

The profile analysis reported in the previous sections sorted the sample into three groups: Profiles 1, 2, and 3. In this section, we describe each profile using the six teacher characteristics considered in the study: (1) gender, (2) subject preference (between math and statistics), (3) degree level, (4) academic background, (5) teaching experience, and (6) four dimensions of teachers' desire to improve their teaching: (a) self-awareness of the need for more content knowledge (CK), (b) self-awareness of the need for more pedagogical content knowledge (PCK), (c) willingness to participate in professional development (PD) programs to improve CK, and (d) willingness to participate in PD programs to improve PCK. The first five characteristics along with the four-part sixth characteristic give a total of 11 variables. We test the independence between each variable and the three-way grouping using Pearson's chi-squared tests. To determine to what extent the group differences could be explained by each of the 11 independent variables, we set the following hypotheses for each variable i, where $i = 1, 2, 3, 4, 5, 6a, 6b, 6c,$ and $6d$:

- Null hypothesis (H_0): the group difference is not explained by variable i.
- Alternative hypothesis (H_1): the group difference is explained by variable i.

In checking these differences, we considered the effect sizes using the scores of Cramer's V, which is known to be well-suited for checking hypotheses of chi-square distribution. In this study, we classify Cramer's V scores as *very strong* ($> .25$), *strong* ($> .15$), *moderate* ($> .10$), *weak* ($> .05$) and *non-existent or very weak* (> 0) (Cramer, 1946).

Gender

Participants were asked, "What is your gender?" with listed choices of "male," "female," and "not applicable." Outcomes of the chi-square test are summarized in Table 6. The last column of the table shows that the chi-square statistic for "gender" equals 11.610^{**} (significant at $\alpha = .01$) with $df = 2$, and that the effect size, measured by Cramer's V, is .207 (*strong*), which falls into the category of *strong*. The gender difference seems to be most overtly present in P1, and P3 consists only of males.

Table 6. Chi-Squared Distribution of Gender in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
Gender	Male	24 (49.0)	**136 (65.1)**	**12 (100)**	11.610^{**}(2)	.207 (*strong*)
	Female	**25 (51.0)**	73 (34.9)	0 (0)		
	Total	49 (100)	209 (100)	12 (100)		

Subject Preference

Regarding the subject preference for teaching between mathematics and statistics, participants were asked, "Which subject do you enjoy teaching more?" with listed choices of "mathematics," "statistics," "no preference," and "N/A." Outcomes of the chi-square test are summarized in Table 7. The last column of the table shows that the chi-square statistic for "preference" equals 23.601^{***} (significant at $\alpha = .001$) with $df = 4$, and that the effect size, measured by Cramer's V, is .209 (*strong*). A difference seems to be overtly present between P1 and P2.

Table 7. Chi-Square Distribution of Statistic for Subject Preference in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
Subject preference	Math	**32 (66.7)**	63 (30.1)	**6 (50.0)**	23.601^{***}(2)	.209 (*strong*)
	Stats	8 (16.7)	**74 (35.4)**	2 (16.7)		
	No pref.	8 (16.7)	72 (34.4)	4 (33.3)		
	Total	48 (100)	209 (100)	12 (100)		

Degree Level

Participants were asked, "What is the highest degree you hold?" with listed choices as follows:

- Bachelor's (or equivalent), and currently not enrolled in a graduate degree program;
- Bachelor's (or equivalent), and currently enrolled in a graduate degree program;
- Master's (or equivalent), and currently not enrolled in a doctoral degree program;
- Master's (or equivalent), and currently enrolled full-time in a graduate degree program; and
- Doctoral (or equivalent).

Outcomes of the chi-square test are summarized in Table 8. The last column of the table shows that the statistic for *degree level* equals 32.330^{***} (significant at $\alpha = .001$) with $df = 8$, and that the effect size, measured by Cramer's V, is .245, which would fall between the *very strong* and *strong* categories depending on rounding. As the table indicates, no clear degree-level difference was found.

Table 8. Chi-Square Distribution of Degree in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
Degree	Bachelor	1(2.0)	0(0.0)	0(0.0)	32.330^{***}(8)	.245 (*very strong/ strong*)
	Bachelor +	0(0.0)	0(0.0)	1(8.3)		
	Master	9(18.4)	67(32.1)	1(8.3)		
	Master +	4(8.2)	9(4.3)	1(8.3)		
	Doctoral	**35(71.4)**	**133(63.6)**	**9(75.0)**		
	Total	49(100)	209(100)	12(100)		

Academic Background

Regarding the instructors' academic background, we used participants' responses to the item, "What is your highest degree in?" with a listed choice of "Math education (or Math teaching)," "Mathematics," "Statistics (or Mathematics with a Statistics concentration)," and "other." For data analysis, participants were grouped as: (1) Math: having a graduate degree with the highest degree in math without holding a graduate degree in statistics or math education; (2) Stat: having a graduate degree in statistics; and (3) MaEd: having a graduate degree with the highest degree in math education without holding a graduate degree in statistics. Outcomes of the chi-square test are summarized in Table 9. The last column of the table shows that the chi-square statistic for *degree level* equals 11.193* (significant at $\alpha = .05$) with *df* = 4, and that the effect size, measured by Cramer's V, is .145, which would again fall between the *strong* and *moderate* categories depending on rounding. As the table shows, no clear degree-level difference was found.

Table 9. Chi-Square Distribution of Academic Background in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
Academic Background	Math	**39(81.3)**	**132(63.2)**	**8(72.7)**	11.193*(4)	.145 (*strong/ moderate*)
	Stat	2(4.2)	50(23.9)	3(27.3)		
	Math Ed	7(14.6)	27(12.9)	0(0.0)		
	Total	48(100)	209(100)	11(100)		

Statistics Teaching Experience

The participants were asked, "How many years have you taught statistics?" They were directed to mark a number on a bar that ranged from 0 to 50. To ease data analysis, we grouped the participants into three groups: (1) novice: 5 years or less of experience; (2) mid-exp: 6–15 years of experience; and (3) high-exp: 16 years of experience or more. Outcomes of a chi-square test are summarized in Table 10. The last column of the table shows that the chi-square statistic for *experience* equals 9.785* (significant at $\alpha = .05$) with *df* = 4, and that the effect size, measured by Cramer's V, is .135 (*moderate*). Note that P1 (57.1%) contains more novice instructors than P2 (33.5%) or P3 (41.7%).

Table 10. Chi-Square Distribution of Experience in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
Experience	Novice	**28(57.1)**	70(33.5)	**5(41.7)**	9.785*(4)	.135 (*moderate*)
	Mid-exp	11(22.4)	**83(39.7)**	4(33.3)		
	High-exp	10(20.4)	56(26.8)	3(25.0)		
	Total	49(100)	209(100)	12(100)		

Desire to Improve Teaching

The participants were asked to respond to four statements: (a) "I need to have more statistics content knowledge to be able to teach elementary statistics more effectively"; (b) "I need to have more pedagogical content knowledge for statistics to be able to teach elementary statistics more effectively"; (c) "I would like to participate in professional development programs to learn more statistics content"; and (d) "I would like to participate in professional development programs to gain more pedagogical content knowledge of statistics." While these items were rated on a 7-point Likert scale, we combined the scores into three groups to enable chi-square tests: Group 1 for scores 1–2, Group 2 for scores 3–5, and Group 3 for scores 6–7. Chi-square tests were conducted separately for each of the four statements, but the outcomes of all four are summarized in Table 11.

The outcomes of the chi-square test regarding gaining statistics content knowledge are summarized in row (a) of Table 11. The first cell of the last column shows that the chi-square statistic equals 10.658* (significant at α = .05) with *df* = 4, and that the effect size, measured by Cramer's V, is .140 (*moderate*). The measures in row (a) of the table indicate that individuals in P1 (30.6%) feel more need to gain statistics content knowledge than those in P2 (15.3%) or P3 (25.1%), and that most of P3 (66.7%) sees little to no need to do so. The outcomes of the chi-square test regarding gaining pedagogical content knowledge for statistics are summarized in row (b). The second cell of the last column shows that the chi-square statistic equals 4.118 (not significant at α = .05) with *df* = 4, and that the effect size, measured by Cramer's V, is .087 (*weak*). The measures in row (b) indicate that individuals in P1 (28.6%) feel more need to gain pedagogical content knowledge than those in P2 (16.3%) or P3 (16.7%), and that many of P2 (41.1%) sees little to no need to do so.

Table 11. Chi-Square Distribution of Four Sub-variables of Desire to Improve Teaching in Each Latent Profile

Variables \ Profile		P1 (%)	P2 (%)	P3 (%)	X^2 (df)	Cramer's V
(a)	1	**17(34.7)**	**108(51.7)**	**8(66.7)**	10.658*(4)	.140 (*moderate*)
	2	**17(34.7)**	69(33.0)	1(8.3)		
	3	15(30.6)	32(15.3)	3(25.0)		
	Total	49(100)	209(100)	12(100)		
(b)	1	16(32.7)	**86(41.1)**	4(33.3)	4.118 (4)	.087 (*weak*)
	2	19(38.8)	88(42.1)	6(50.0)		
	3	**14(28.6)**	**35(16.7)**	**2(16.7)**		
	Total	49(100)	209(100)	12(100)		
(c)	1	10(20.4)	68(32.5)	**7(58.3)**	9.037 (4)	.129 (*moderate*)
	2	**27(55.1)**	**83(39.7)**	2(16.7)		
	3	12(24.5)	58(27.8)	3(25.0)		
	Total	49(100)	209(100)	12(100)		
(d)	1	9(18.4)	50(23.9)	**8(66.7)**	13.544*(4)	.158 (*strong*)
	2	**27(55.1)**	**94(45.0)**	2(16.7)		
	3	13(26.5)	65(31.1)	2(16.7)		
	Total	49(100)	209(100)	12(100)		

The outcomes of the chi-square test regarding the desire to participate in PD programs to learn more statistics content are summarized in row (c). The third cell of the last column shows that the chi-square statistic equals 9.0374* (not significant at α = .05) with *df* = 4, and that the effect size, measured by Cramer's V, is .129 (*moderate*). The measures in row (c) indicate that individuals in P3 have less desire to participate in PD programs to learn more statistics content knowledge than their counterparts in P1 or P2.

Finally, the outcomes of the chi-square test regarding the desire to participate in PD programs to learn more pedagogical content knowledge are summarized in row (d). The fourth cell of the last column shows that the chi-square statistic equals 13.544* (significant at α = .05) with *df* = 4, and that the effect size, measured by Cramer's V, is .158 (*strong*). The measures in row (d) indicate that individuals in P3 have less desire to participate in PD programs to learn more statistics pedagogical content knowledge than their counterparts in P1 or P2.

Discussion

The goals of this study are: (1) to identify distinct groups within the population of college instructors of statistics regarding their attitudes toward teaching CIs in their elementary statistics courses and toward using CIs for inferential statistics, and (2) to explain the heterogeneity among the distinct groups in terms of a set of teacher characteristics. Regarding the first research goal, the latent profile analysis suggested a classification of the population of college instructors of statistics into three profiles based on their attitudes toward teaching and using confidence intervals. Using the instructors' responses to the survey items, we were able to describe the three profiles as a "stable-score group," a "high-score group," and a "diverse-score group." In addition, under the assumption that there are three subgroups, we found four items (Items 2, 3, 4, and 6) for which the most significant differences were shown among the three profiles. It is interesting to note that all of these four items are from the *independent* view; that is, the view in which CIs are considered without regard to *p*-values. A possible reason for this is that the in-comparison view may promote dichotomous thinking, as the items in this view may encourage the respondent to compare, for example, the benefits of CIs with the benefits of *p*-values. We can reasonably assume that many instructors hold more positive attitudes toward *p*-values than confidence intervals due to greater exposure to and familiarity with inferential statistics that uses *p*-values, as learners, practitioners, and educators. While participants are likely to show various ranges of feelings, beliefs, and behaviors regarding confidence intervals, these ranges might be reduced by item statements confined to comparison between the use of CIs and the use of *p*-values. Note also that three of the four items (Items 2, 3, and 4) are from the teaching context. This may mean that teachers in different profiles show greater differences in the teaching than in the general context. That is, those in the different profiles can be distinguished more by their attitudes toward teaching the topic of confidence intervals than by their use of confidence intervals for inferential statistics. In addition to the findings related to the first research question, we note a few other interesting findings: (1) the scores are low for Items 9 and 11 for all three profiles, and the group differences were not significant at the significance level of $\alpha = .05$; both of these items are from the in-comparison context as opposed to the independent context; (2) on each of 10 out of the 12 items, Profile 2's scores are significantly

higher than the scores on the same items of at least one of the other two profiles based on the *F*-values; and (3) Profile 3 scored significantly higher than Profile 1 on Items 2, 3, 6, 8, 10, and 12, four of which address behavior.

Regarding the second research goal, the chi-square tests showed that the profile differences may be explained by some of the six biodata variables (i.e., 11 variables with the four subitems of the sixth variable). The results indicate that P1 (the stable-score group) in comparison to P2 or P3, has more females, more individuals who prefer math, and more who hold terminal degrees, as well as significantly fewer individuals with a statistics background, who have less experience in statistics teaching and more motivation to learn content or pedagogical content knowledge. P2, in comparison to P1 and P3, has more males and more individuals who have master's degrees, prefer statistics over math, and have more experience in statistics teaching, as well as less motivation to learn content or pedagogical content knowledge.

This study has limitations. Although the sample size (n = 268–270) allowed quantitative data analysis for statistical inferences, a larger scale study is necessary to arrive at firmer conclusions. Furthermore, the voluntary responses in the data collection, and the low response rate (2.67%) constrain the extent to which the findings can be generalized. Notwithstanding these limitations, the findings of this study have implications for the teaching and learning of elementary statistics at the college level. With the advantages that confidence intervals (CIs) provide in inferential statistics, scholars of alternative statistical methods encourage wider use of CIs to the research community of social sciences (Cumming & Finch, 2001) and suggest rules of understanding CIs as an alternative statistical method to use in inferential statistics (Cuming & Finch, 2005). If such recommendations to use CIs to a wider extent are to be successfully implemented in the research community, it will be important for college instructors of statistics to be familiar with the concept of CIs, the use of CIs, and the teaching of the topic. Therefore, CIs should be considered as essential as *p*-values as a part of inferential statistics at the college level of elementary statistics (Fidler and Cumming, 2005). The authors of this paper conducted the study that has taken one of the first steps in this direction, by identifying different types of instructors in its three profiles. This profiling allowed us to learn the characteristics of instructor groups that showed differences in the ways they develop attitudes toward using and teaching confidence intervals, and to identify the kind of instructor group that needs more attention to develop more positive attitudes toward teaching and using CIs. It is hoped that the study lays a foundation for future research exploring the origins and factors of ESIs' attitudes toward CIs, the dynamics of how ESIs' attitudes affect their teaching and pedagogical decision-making, and how the challenges in teaching CIs relate to elementary statistics students' attitudes.

References

Aiken, L. R., Jr. (1970). Affective factors in mathematics learning: Comments on a paper by Neale and a plan for research. *Journal for Research in Mathematics Education, 1*(4), 251–255.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7*(2), 20–39.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education, 17*(2). [Retrieved from: www.amstat.org/publications/jse/v17n2/castrosotos.html]

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530–572.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170–180.

Estrada, A. (2002). *Análisis de las actitudes y conocimientos estadísticos elementales en la formación del profesorado* [Analysis of attitudes and elementary statistical knowledge in training teachers] (Unpublished doctoral dissertation). Universidad Autónoma de Barcelona.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* (Unpublished PhD thesis). University of Melbourne.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie/Journal of Psychology, 217*(1), 27–37.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2). [Retrieved from http://www.amstat.org/publications/jse/v2n2/gal.html]

Geiser, C. (2012). *Data analysis with Mplus*. New York: Guilford Press.

Gelman, A. (2013). *P* values and statistical practice. *Epidemiology, 24*(1), 69–72.

Gómez-Chacón, I. (2000). Affective influences in the knowledge of mathematics. *Educational Studies in Mathematics, 43*(2), 149–168.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*(2), 215–231.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164.

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88.

Kim, H., Wang, X., Lee, B., & Castillo, A. (2017). College instructors' attitudes toward statistics. In A. Chronaki (Ed.), *Proceedings of the Ninth International Mathematics Education and Society Conference* (Vol. 2, pp. 611–621). Volos, Greece: University of Thessaly.

Kim, H., Wilson, A. T., & Ko, H. (2018). College instructors' attitudes toward confidence intervals. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of ICOTS 10, Kyoto, Japan*. Voorburg, The Netherlands: International Statistical Institute.

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of ICOTS8, Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Lindley, D. V. (1999). Comment on Bayarri and Berger. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. 6, p. 75). Oxford, UK: Clarendon.

Marden, J. I. (2000). Hypothesis testing: From *p* values to Bayes factors. *Journal of the American Statistical Association, 95*, 1316–1320.

Martins, J. A., Nascimento, M. M. S., & Estrada, A. (2012). Looking back over their shoulders: A qualitative analysis of Portuguese teachers' attitudes towards statistics. *Statistics Education Research Journal, 11*(2), 26–44.

McLachlan, G., & Peel, D. (2000). Finite mixture models. New York, NY: John Wiley.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nelder, J. A. (1999). From statistics to statistical science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*(2), 257-269.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569.

Olson, J. M., & Zanna, M. P. (1993). Attitude and attitude change. *Annual Review of Psychology, 44*, 117–154.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *12*(1), 103–124.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht, The Netherlands: D. Reidel.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Hillsdale, NJ: Erlbaum.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343.

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling, 20*(4), 640-657.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*(5), 508–518.

Appendix. Survey Items in Part II

1. I like reading newspaper or scholarly articles that used CIs for making statistical inferences.
2. I rarely explain CIs in my classes as a statistical tool for making decisions about data.
3. I feel uncomfortable teaching the underlying concept of CIs.
4. CIs should be discussed as a method for making statistical inferences in an introductory statistics course curriculum.
5. CIs provide detailed information on how inferential conclusions are reached from data.
6. I rarely approach inferential statistics using CIs.
7. It is important to convince students that CIs are as valid a tool as *p*-values for statistical inference.
8. I find decisions about data less clear when they rely on CIs as opposed to *p*-values.
9. I feel more comfortable teaching CIs than *p*-values.
10. In introductory statistics classes, I emphasize CIs as a tool for inferential statistics as much as *p*-value methods.
11. I tend to treat statistical inference problems in terms of CIs.
12. CIs provide less evidence for making statistical inferences than the *p*-value methods.