EVALUATION OF THEORETICAL AND EMPIRICAL CHARACTERISTICS OF THE COMMUNICATION, LANGUAGE, AND STATISTICS SURVEY (CLASS)

AMY E. WAGLER
The University of Texas at El Paso
awagler2@utep.edu

LAWRENCE M. LESSER
The University of Texas at El Paso
Lesser@utep.edu

ABSTRACT

The interaction between language and the learning of statistical concepts has been receiving increased attention. The Communication, Language, And Statistics Survey (CLASS) was developed in response to the need to focus on dynamics of language in light of the culturally and linguistically diverse environments of introductory statistics classrooms. This manuscript presents a cross-cultural evaluation of the characteristics of the CLASS III (third generation of the instrument) and provides a user-friendly cross-culturally valid version of the CLASS. Mixed methods are employed to investigate further characteristics of the CLASS III and provide a scale (CLASS IV) that may be utilized in diverse settings. These research results have implications for instructors, professional developers, and researchers to improve instruction with culturally and linguistically diverse student populations.

Keywords: *Statistics education research; Culturally and linguistically diverse student; Cross-cultural communication; Teacher education; English language learner*

1. INTRODUCTION

In recent years, issues of language and how the acquisition of statistical vocabulary can affect conceptual understanding have received considerable attention in the statistics education literature. Some researchers have explored how introductory statistics students acquire knowledge of statistical terms and how use of particular statistics terminologies can inhibit or promote statistical knowledge (Dunn, Carey, Richardson, & McDonald, 2016; Kaplan, Fisher, & Rogness, 2009, 2010; Whitaker, Jacobbe, & Foti, 2014). Another emerging focus in statistics education research is English Language Learners (ELLs) and how interactions of everyday and academic registers affect an ELL introductory student's learning of statistical concepts. With this focus, this study follows up on a qualitative case study (Lesser & Winsor, 2009) and a quantitative exploratory study (Lesser, Wagler, Esquinca, & Valenzuela, 2013) in order to formally examine the validity of the third generation of the Communication, Language, And Statistics Survey (CLASS III) and to revise the CLASS to provide a streamlined version (CLASS IV). In order to refine the CLASS III instrument further, the scale is assessed for characteristics of cross-cultural equivalence (Hui & Tirandis, 1985). This study entails a mixed methods analysis of the characteristics of the CLASS III items to determine how well they assess the theoretical constructs upon which the CLASS was developed. Moreover, special attention is paid to whether concepts of the CLASS III transfer well between cultures and are not culture specific. Establishing the cultural relevancy of the CLASS in settings with other student populations will allow researchers to understand the cultural and linguistic factors that affect learning in the statistics classroom. The qualitative and quantitative aspects of this study are

integrated in this manuscript to provide the statistics education research community with a streamlined scale that may be used in varied research settings.

In this manuscript, we utilize the term ELL for English Language Learner and non-ELL for someone who has the fluency of a native English speaker. This distinction is important because a growing proportion of college and university students in the United States has or is acquiring English as a second, third, or even fourth language (Payán & Nettles, 2008) and this trend is continuing in most regions of the U.S. Moreover, past research (Lesser & Winsor, 2009; Lesser et al., 2013) has indicated that those acquiring English experience particular and distinctive differences when learning introductory statistical concepts and vocabulary. Issues of equity and diversity compel statistics education professionals to recognize these trends and revise our teaching and curriculum to reflect these trends (Lesser, 2010).

CLASS III, the current version of the CLASS, is intended to be utilized by researchers needing to distinguish among the cultural and linguistic factors that affect learning in introductory statistics. Construction of the CLASS started with a qualitative study (Lesser & Winsor, 2009) and development of a pilot scale followed. A comprehensive description of the scale components and development is contained in Section 2.1. In order for the CLASS III to be employed in this manner, it is necessary that researchers be assured that observed differences between student groups are measuring salient factors (i.e., those related to the cultural and linguistic factors involved in learning statistical concepts). Assuring this includes showing that the factors are not dependent primarily on measurement artifacts unrelated to these cultural or linguistic influences. For example, culturally-based response tendencies can bias item scores in ways that make differences among respondents appear salient when they are actually just cultural artifacts. These issues will be discussed further in Section 2. In addition to showing the CLASS to be valid across cultures, the CLASS III must be made more usable to researchers and practitioners. The length of the scale needs to be shortened so that users can readily obtain a sample large enough to yield a reasonable participant-to-item ratio and the sample independent item and test characteristics can be made available for future users of the scale. The full version of the CLASS III is available in Appendix A and the revised CLASS IV may be requested from the first author.

In short, this manuscript seeks to demonstrate that an identified subset of CLASS III items is functional and shows evidence of cross-cultural equivalence in the tradition of Hui and Tirandis (1985). The goal of this study is to provide a scale that is 1) useful for recognizing learning and teaching preferences among students with varying cultural and linguistic backgrounds, 2) able to assess cultural and linguistic-based differences students experience when learning introductory statistics, and 3) able to identify when and where interactions between the everyday and academic registers provide help or present difficulty for culturally and linguistically diverse students.

2. RESEARCH ON CULTURAL/LINGUISTIC BACKGROUND IN STATISTICS

In the statistics education community, a sustained research focus has recently developed regarding language-based concerns in learning introductory statistical concepts. This research article builds on this literature and further expands the research base by grounding the study in the linguistic theory of register (Halliday, 1978). In particular, many researchers have explored how language use affects learning in introductory statistics and probability (Green, 1984; Kaplan, Fisher, & Rogness, 2009, 2010; Kaplan, Rogness, & Fisher, 2011, 2014; Lavy & Mashiach-Eizenberg, 2009; Richardson, Dunn, Carey, & McDonald, 2016; Whitaker, 2016). All of these research studies involve English as the native language and focus on the lexical ambiguity of technical terms commonly used in statistics and probability. Hubbard (1991) does address ELLs in the statistics classroom, but does not explore the issues in a research-based manner. Other studies have focused on language learners with an emphasis on learning probability concepts (Kazima, 2006; Phillip & Wright, 1977) and are also expository rather than empirical. Appendix B outlines factors that may affect the validity of the CLASS for a general audience and may be skipped for those familiar with these issues.

2.1. CONSTRUCTION OF THE CLASS

The construction of the CLASS I instrument began in the summer of 2009 when the second author and two other collaborators met to design a pilot survey. The items of this pilot survey were informed by the qualitative research study on ELLs in introductory statistics (Lesser & Winsor, 2009). The themes emerging from the qualitative study included deficiencies in cognitive academic language proficiency (CALP), misunderstanding of the context, student practices and beliefs, preferences with regard to teaching strategies, and transfer between different CALPs (i.e., academic subjects with overlapping terms). These emergent themes demonstrated the different interactions with language as theorized by the theory of register (Halliday, 1978). Following a 2009 qualitative study, a 2013 quantitative analysis of the items proceeded (Lesser et al., 2013). The 2009 study focused on describing the characteristics of ELL and non-ELL populations when learning introductory statistics. The themes of Lesser and Winsor (2009) informed some revisions of items from CLASS I to form an updated version of the CLASS, denoted CLASS II. Additionally, the themes emerging from the qualitative study were updated with a more modern theory of language—register theory. Register theory (Halliday, 1978) purports that a *register* (i.e., a variety of language used in a specific situation) can be described by three attributes: *field* (the subject matter of the communication), *mode* (the type of communication, such as oral or written), and *tenor* (the social relationships involved in the communication). For more details about the theory of register and analysis of the field, mode, and tenor item sets, see Lesser et al. (2013). We also updated the CLASS II with some wording and formatting changes in order to make the scale easier to read for future use. This new version with the minor wording and formatting changes is denoted CLASS III, which is the version of the scale utilized for the data collection and analysis included here. With the CLASS III, a response of 0 was also allowed with the label "I don't understand" to decrease the incidence of non-meaningful responses caused by incomprehension of what is stated in the item, which was noted by respondents in past administrations of the CLASS.

In addition to items addressing the dimensions of register, some textbook questions were added to the CLASS I instrument in order to assess any differences in how ELLs versus non-ELLs interpret these questions and questions regarding the student's background were also included in the CLASS I. For thorough analysis of these concepts, see the original study (Lesser & Winsor, 2009). Henceforth, the set of items constructed directly following the qualitative study will be called CLASS I. Table 1 summarizes the development of the CLASS over time, beginning with the CLASS I.

Table 1. Chronology of CLASS instrument development

Version	Year	Changes	How it was used
I	2009	Original version	Valenzuela (2009)
II	2010	Wording changes	Lesser, Wagler, Esquinca, & Valenzuela (2013)
III	2011	Added "I don't understand" option and reformatted scale for ease of reading	To better understand structure and dimensionality of the scale and allow students to voice confusion with items
IV	2015	Revised version after cross-cultural equivalency assessed	To improve accessibility and usefulness for cross-cultural populations

A preliminary analysis of the data is described in Wagler and Lesser (2014). This study considered just the dimensionality of the CLASS III for ELL and non-ELL populations. Modified parallel analysis, factor analysis, and reliability analysis assess the construct validity of the scale for both populations. The analysis in Wagler and Lesser identified three subscales for the field dimension of register, two subscales for the mode dimension of register, and one subscale for the tenor dimension of register. The results of this study are summarized in Table 2. Items 1–3 and 36–39 (see Appendix A for items) ask questions about the student's background and assess whether the student's native language is English. These items are utilized to describe the ELL and non-ELL populations. Note that this study evaluates only how the

items co-vary in order to assess the dimensionality of the scale. Some of the items that did not align well to the theoretical construct are still considered for use in the CLASS IV in later sections. For more details about the empirical evidence for the dimensionality of the CLASS III, see Wagler and Lesser (2014). In the next section we provide additional analysis in order to assess the nomological properties (e.g., how well the CLASS III conforms to register theory and discriminates among populations) of the CLASS III and to examine empirically the item characteristics.

Table 2. Register subscales in the CLASS III instrument

Subscale	Items for both ELLs and non-ELLs	Items for ELLs only
Field		
Word Confusion	30, 31, 32, 34	43, 48, 49
Technical Words	15, 16	
Everyday Words	4, 6, 21, 22, 29, 35	
Mode		
Modes of Words	13, 19, 23, 24	40, 41, 42, 44, 45,
Negative Wording	26, 27, 28	46, 47
Tenor	14, 18, 36	no items

The items of the CLASS III instrument were categorized individually into the appropriate subscales. After meeting and comparing results, we compared individual results and found that 92% of the items were classified identically, leaving four items with different categorizations. The authors discussed rationale for their categorizations and then re-classified the items, reaching 100% agreement after this iteration. The CLASS III items are derived from the CLASS II, and the item order is changed so that the last few items of the CLASS III apply only to ELLs (see Appendix A for detailed information). The CLASS III instrument appears as the Appendix of Lesser et al. (2013).

2.2. EQUIVALENCY FOR CROSS-CULTURAL POPULATIONS

There is a need for more studies focusing on how cross-cultural populations approach learning in statistics. To meet this goal, scales used need to show evidence of cross-cultural equivalency. This is a three-level process that incorporates qualitative and quantitative analysis of the scale (see Appendix B for details).

The subscales of the CLASS III discussed in Section 2.1 are measured by a set of items designed to reflect the emerging themes described in Lesser and Winsor (2009) and constructed in light of the theory of register (Halliday & Matthiessen, 2004). In order to assess the fidelity of each item to the theoretical construct of register and to assess how well these items align to the dimensions of register theory, a combination of arguments from a theory basis and item response modeling is employed. The theory will inform the construct and functional equivalence of the CLASS while item response modeling will provide evidence about the operationalization, item, and scalar equivalence of the CLASS.

Level One: Conceptual equivalence of the CLASS III As outlined in Lesser et al. (2013), the theory of register underlies the construction of the CLASS III. Because Lesser et al. provides an overview about the theory of register, we now focus on the generalizability of the theory of register across people of various cultural and linguistic backgrounds. Pepitone and Triandis (1987) advocated that three characteristics must hold for a theory to be cross-culturally generalizable. First, the theory must be valid across social behavior of different form and content. This holds for register theory as it describes how language varies according to content (field), mode (practice), and tenor (also an aspect of practice). A second characteristic necessary for establishing generalizability of a theory is that the theory must hold "across different situational contexts" (Pepitone & Triandis, 1987). Register theory describes variations in language used in any setting: from academic to purely social settings. Thus, this characteristic holds as

well. Finally, a theory must also hold for diverse populations of people. Halliday's theory of register has been used to describe varieties of language use in many different cultures and language groups (Liu, 2014). This criterion also holds and thus the cultural equivalence of register theory appears valid.

Thus, the concept of register may be understood as a mediator between a culture/setting and how language gets used. A specific culture or context may give rise to a particular register which, in turn, informs the way a student understands communication or communicates. Thus, register theory provides a unifying framework for assessing how language practices vary even in the setting of a statistics learning environment. We note that demonstrating conceptual equivalence of the theory of register in no way implies that the way register gets realized in a particular communication is equivalent. A student from a particular cultural or language background accesses and interprets social cues differently than a student from another background. Analyzing how their communication differs is unified by the theory of register and assists researchers in understanding how students construe meaning based on the field, mode, and tenor dimensions of register.

Complementing the theoretical approach outlined above, two focus group sessions were conducted to assess the CLASS III items for conceptual equivalence with special focus on the relevance of a construct across cultures. Care was taken to include professionals with backgrounds in statistics or multicultural education who personally had varied linguistic and cultural backgrounds (combined focus group members included two Mexican nationals, a national of a country in southern Asia, and three Mexican-American). Focus group participants were invited via email to a one-hour focus group session. All participants were volunteers and were not compensated monetarily, but told they would be acknowledged in the paper and would be provided an opportunity even to co-author the paper (no one exercised the latter option).

The two focus group sessions were structured similarly. The first focus group, held early fall 2013, was provided the complete CLASS III, given time to review the set of items, and provided two prompting questions for each item: 1) "What do you think this item is getting at?" and 2) "Do you see any possible problems or biases with this item (either the idea of the item or the way it is worded), especially with respect to matters of cultural or linguistic background of any university student? If so, feel free to mark suggested changes on the sheet." The second focus group, held early spring 2014, used the same prompts for a subset of CLASS III items that had been found problematic based on feedback from the first focus group. Following time to review the set of CLASS III items, participants were provided an opportunity to ask questions and provide feedback about any particular items. The participants wrote down their comments voluntarily and handed in the forms. Table 3 reports the feedback obtained as a result of the two focus group sessions held during the 2013–14 school year. In addition to the problematic items appearing in Table 3, these focus groups also pointed out items that were either redundant, not relevant, or too difficult to understand. In particular, items 5 and 10 were redundant with respect to other real-world themed items, items 7–9 and 12 were not considered relevant to the CLASS because they failed to conform well to register theory, and items 17, 20, 23, and 25 were too difficult to understand. In particular, items 17 and 25 had more complex sentence structures that made them harder to read and all dealt with topics that are important to instructors but less salient to students (e.g., cultural background, going beyond the definition of a term, wording of a test item not being directly tested).

Level Two: Construct operationalization of the CLASS III Evidence regarding the construct operationalization of the CLASS III (based on the theory of register) was presented in Wagler and Lesser (2014). In this article, we analyze CLASS III items using confirmatory factor analytic models and measures of internal consistency. The analysis showed that a subset of items from the CLASS III was invariant across ELL and non-ELL student populations, and the internal consistency estimates (Cronbach's alpha) were commensurate across these two groups. The analysis confirmed that the structure of the CLASS III aligned to the theoretical construct of register and its dimensions of field, mode, and tenor. The analysis further implied that additional item revision and deletion were necessary. See Appendix C for details about the analysis. This study provided evidence that using this

Table 3. Summary of and response to focus group feedback

Item discussed	Issue raised	Researchers' response
6. Connections to words used in everyday conversation are most helpful to me when I encounter them before I encounter the technical academic terms.	Ambiguous and needing revision or elimination	Wording changed to: "It is helpful when the instructor introduces a new technical word by first talking about how that same word is used in everyday speech."
10. In-class discussion of examples of statistics in the newspaper or a newsmagazine helps me understand statistics concepts.	Students may not be interested in reading news articles.	The item is focused on the instructor bringing in the real-life context and does not require the student to read the news article on her own. Thus, we still consider this item for inclusion in the CLASS IV.
14. Professors often do not wait enough time after asking a question for me to think about what the question means, and think of an answer.	May make respondents hesitate to answer honestly as it appears to criticize the professor	Wording changed to be more focused on the respondent and their perceptions of class wait time. Revision: "Professors often do not wait enough time after asking a question for me to come up with an answer."
26. The phrase "not all group averages are equal" is difficult for me to understand.	Ambiguous because the phrase "not all group averages are equal" might be interpreted by respondents in at least two ways, when only one way is correct	This feedback is disregarded as the intent of the item is to assess whether students understand the phrase. However, upon reflection the authors realized that respondents will often believe they understand the phrase "not all group averages are equal" when in fact they do not. Therefore, the interpretation of the responses is not clear. We considered not using a particular example, but instead asking more generally about the use of negation words. However, this is also problematic because most students are not aware enough of how these linguistic patterns are described. Thus, this item, and the other similarly-worded items, will be revised to reflect more precise wording that does not allow the ambiguity in the response observed in the original forms. The final version used was substantively different and read "Non-statistical words in questions can make it difficult to answer."
36. If I don't understand what is going on in class, I will pretend that I understand when the instructor is looking towards me.	Respondents may not want to answer honestly.	Perhaps this is due to the word "pretend" in the item, which has negative connotations. Revision: "If I don't understand what is going on in class, I will try to appear that I understand when the instructor is looking towards me."

identified subset of items from the CLASS III as a basis for the CLASS IV revision was a reasonable starting point and provided strong evidence of construct operationalization even when the CLASS IV was generalized to other ELL student populations.

Level Three: Item equivalence and scalar equivalence of the CLASS III Evidence regarding the item and scalar equivalence of the CLASS III is presented in the remainder of this paper using two-parameter polytomous item response models and comparison of Item Characteristic Curve (ICC) differences across the research groups. This analysis provides evidence specifically about the item and scalar equivalence about the subset of CLASS III items identified as having construct operationalization in level two and establishes the likelihood of equivalence for the CLASS IV, which

is formed as a result of the theoretical and empirical evidence presented in this manuscript. In the results section, we refer to evidence concerning item equivalence by the label 3a and denote evidence concerning scalar equivalence as 3b.

3. METHOD

3.1. SETTING AND PARTICIPANTS

This research study took place at a moderately large doctoral/research university and community college system located in an urban setting in the Southwestern United States. In this urban region, 82% of residents are Hispanic and 71% of families with school-age children report Spanish as the preferred language at home. At the research university, roughly 80% of the student population is Hispanic and about 5–10% of the Hispanic student population are Mexican nationals who commute across the border to take courses. During the course of the research study, the proportion of students required to take the Test of English as a Foreign Language (TOEFL) ranged from 6.3% to 6.9%, according to the university's center for institutional data analysis. There is a fairly specialized criterion for being required to take the TOEFL—namely, the student must have all prior degrees from a non-English speaking country. This excludes any U.S. citizens and permanent residents that do not speak English as their dominant language. Approximately 40% of students at the authors' university self-identify as non-native English speakers, thus providing a critical population of potential ELL students. Moreover, it is known that 49% of students entering this university need remedial coursework focused on reading, writing, and mathematics (FSG Social Impact Consultants, 2011). Some of the survey respondents attend a regional community college system also in this urban city in the Southwestern United States. The representation of Hispanics in the community college system, which is 88%, exceeds the overall proportion of Hispanics in the city. FSG Social Impact Consultants (2011) notes that 63% of students need remedial coursework in reading, writing, and mathematics upon entering the community college system. The makeup of this population should be similar to the university population with perhaps more representation of students with limited English proficiency.

The participants consisted of all students attending one class meeting in the second and third weeks of November of the introductory statistics course in its five fall 2011 sections (each consisting of an instructor and about 20–40 students) offered at the university and at the community college system previously mentioned. The courses are described as statistical literacy courses. Students were not offered compensation for the survey, which took 20 minutes of class time to complete. All students in attendance that first day of class agreed to participate (with no one withdrawing later). This course is required only for pre-service elementary and middle school teachers, and the vast majority of the students who enrolled are pre-service elementary teachers, with some pre-service middle school teachers, and a small number of non-education majors taking the course as a way to satisfy core curriculum requirements of the university. A large proportion of the students in these sections were female, mirroring the demographic of the regional population of pre-service elementary teachers. In particular, the official enrollments for the five sections combined were about 87% female.

Of the 560 students taking the full survey, 230 self-identified as speaking a language other than English as their first language (of these, 223 reported a mother tongue of Spanish), 288 self-identified as speaking English as their first language, and 42 did not self-identify either way (and were therefore dropped from the analysis). This left 511 student responses for the analysis using 36 items (an approximate 15:1 student to item ratio). Thus 44.4% of students self-identify as not speaking English as their first language. A question about a student's first language is used in educational settings to identify students who may not be fully proficient in academic English. Although this kind of question may be vulnerable to over-identification of ELLs, it is a useful proxy in many settings. Therefore, we classify the students who identify that English was not their first language as ELLs, while acknowledging that there is a continuum of the degree of "ELL-ness" among this population. Whereas the case study and survey stages of the development of the CLASS did not involve the same students, the high number of ELLs in the survey provides a strong bridge, and a further connection comes from the fact that the modal gender and ethnicity of the survey participants (and of these institutions of higher education) match the gender and ethnicity (Latinas) of the students in the case study of Lesser and Winsor (2009). Multiple imputation is also utilized whenever the data is found to be "missing at

random," using the R package *mirt* (Chambers, 2012). There were 59 observations with at least one missing value and two observations with more than 20 missing values. The two observations were dropped from the analysis and the remaining 57 were retained. For the missing data imputation, a five-factor exploratory IRT model was assumed with quasi-Monte Carlo EM estimation utilized because there were more than three levels per item. However, the 11 respondents found to have dropped out of the survey (e.g., quit responding after a certain item number) are not utilized in the analysis.

Because roughly three-quarters of CLASS III items were for both ELLs and non-ELLs, the researchers were able to administer a single survey discreetly to the entire sample, with a simple instruction for questions after #53 to be answered only by ELLs (six non-ELLs answered these anyway, and those items of their surveys were ignored as they had self-identified as non-ELL).

Hypotheses The following research hypotheses guide the investigation into the equivalence of operationalization of the construct, item, and scalar properties of the CLASS III. Only the items of the CLASS III demonstrating these characteristics of operationalization, item, and scalar equivalence will be retained for a reduced and improved version of the CLASS, denoted CLASS IV. The research hypotheses address the issues of item (3a) and scalar (3b) both assessing level three of cross-cultural equivalence and may be summarized as:

RH1: CLASS IV items exhibit item equivalency across ELL and non-ELL populations.

RH2: CLASS IV items exhibit scalar equivalency across ELL and non-ELL populations.

3.2. ANALYSIS

In this section, we describe item response models appropriate for ordinal (Likert) response data in the context of the validation of the CLASS III. In order to assess the item and scalar equivalency of the CLASS III (Level Three), we analyzed the data using item response theory (IRT) models suitable for ordinal responses. The characteristics of the scale are reported via the item and test parameter estimates from the IRT model.

Multidimensional Item Response Model The analysis focuses on assessing the functionality of CLASS III items and providing the item and test information needed to demonstrate the degree of fidelity to item and scalar equivalence. A multidimensional polytomous IRT model is a useful tool for modeling Likert data when exploring the functionality of items in a scale (Cai, 2010; Chalmers & Flora, 2014) using the *mirt* package in R (Chalmers, 2012). We note that *mirt* package is used for both missing value imputation as described in Section 3.1 and now for modeling the complete data. For each item of a scale, the multidimensional polytomous IRT model predicts the probability of a response for a particular response pattern. The scale allows for responses ranging from 1 to 6 where 1 indicates strong disagreement, 6 indicates strong agreement, and neutral responses are not available (1=Strongly Disagree, 2=Disagree, 3=Somewhat Disagree, 4=Somewhat Agree, 5=Agree, 6=Strongly Agree). The response level 0 will not be included in the analysis and we note that two items (9 and 25) had '0' response rates of 5% and 16% respectively. Additionally, item 35 was the next highest with a 4% rate and all others were less than 1.5%. For this among other reasons, items 5 and 16 are no longer to be considered for inclusion in CLASS IV. We utilized a form of the Multidimensional IRT model, called the Multidimensional Partial Credit model as described by Reckase (2009, p. 106), is given by

$$p(u_{ij} = k \mid \boldsymbol{\theta}_i) = \frac{\exp(\sum_{l=1}^{k} \theta_{il} - b_{il})}{1 + \exp(\sum_{r=1}^{k_i} \theta_{il} - b_{ilr})}$$

for $k = 1, \ldots, k_i$ and where the proportion $p(u_{ij} = k \mid \boldsymbol{\theta}_i)$ is the probability of the j^{th} respondent having response k on the i^{th} item, b_{ilk} is the item difficulty parameter for the i^{th} item, k^{th} response level, and l^{th} dimension, θ_{il} is the latent trait for an l^{th} dimension and i^{th} item. The random variable u_{ij} records the Likert rating (from 1 to k) of the j^{th} respondent on the i^{th} item. This model will allow each item to load onto only a single factor and force the latent traits to be independent.

In the analysis, if the multidimensional polytomous IRT model can accurately predict the response pattern for every item in the instrument, then the items are deemed to conform to the theoretical dimensions proposed in Section 3.2. Fit measures such as the Drasgow, Levine, and Williams (1985) *Z* score method, and the Kang and Chen (2008) modified $S\text{-}X^2$ likelihood ratio test method are computed to assess adequacy of fit and specifically to address the item equivalence of the CLASS III (3a). Reise (1990) suggests that both the *Z* score method (*Zh*) and the modified $S\text{-}X^2$ likelihood ratio test method should be considered when looking for evidence of item misfit. Negative values of *Zh* indicate underfit of the item to the theoretical model and large values of $S\text{-}X^2$ indicate item any type of misfit to the model. In order to avoid type I errors when detecting misfit, we will look for values of *Zh* smaller than -2.998, a *Z* critical point with multiplicity adjustment. In addition to the pointwise significance, tests for item level statistical significance were assessed utilizing the FDR procedure (Benjamini & Hochberg, 1995) and modified Hunter-Worsley (*HW*) procedure (Hunter, 1976; Worsley, 1982). Table 4 shows the results of this item fit analysis. $S\text{-}X^2$ *p*-values are controlled for multiplicity using the false FDR and, additionally, a modified Hunter-Worsley procedure suitable for correlated chi-square distributed endpoints is utilized for comparing each $S\text{-}X^2$ critical point and the FDR adjusted is also applied for comparison. When there is disagreement between the measures, it may be due to the FDR procedure being a little more liberal due to its focus on controlling the false discovery rate, rather than the family-wise type I error rate. In all the analysis, whenever *p*-values are

Table 4. Item fit statistics for CLASS III items
(significant values bolded and indicated beside item number)

	Non-ELL					ELL				
Item	*Zh*	$S\text{-}X^2$	*df*	*HW*(*p*)	*FDR*(*p*)	*Zh*	$S\text{-}X^2$	*df*	*HW*(*p*)	*FDR*(*p*)
4[a]	-1.16	155.6	106	0.00	0.00	-0.97	190.5	125	0.00	0.00
6[a]	-2.27	142.3	108	0.01	0.02	-1.60	148.4	135	0.05	0.22
11	-2.85	143.9	121	0.09	0.16	-2.60	141.2	126	0.17	0.22
13	-2.01	87.4	100	0.67	0.69	-1.60	114.5	89	0.07	0.15
14[b]	-4.65	163.8	153	0.10	0.17	-4.61	154.8	168	0.11	0.53
15[ab]	-6.43	194.3	143	0.00	0.00	-5.11	174.4	135	0.34	0.01
16[b]	-6.03	176.6	151	0.00	0.01	-5.29	168.4	161	0.09	0.15
18[b]	-5.47	152.4	154	0.19	0.24	-5.56	168.8	164	0.05	0.16
19	-1.35	154.9	130	0.12	0.17	-0.90	156.7	125	0.00	0.22
21[ab]	-5.05	166.9	130	0.00	0.01	-4.68	206.2	112	0.42	0.00
22[ab]	-9.58	169.5	98	0.00	0.00	-8.69	199.5	93	0.57	0.00
24[b]	-3.34	114.4	125	0.65	0.69	-2.92	128.6	133	0.00	0.38
26[b]	-7.56	129.4	154	0.24	0.30	-6.32	112.4	142	0.04	0.53
27[ab]	-9.14	154.9	141	0.01	0.02	-9.01	167.9	157	0.00	0.02
28[b]	-11.00	148.4	159	0.12	0.17	-8.64	148.2	143	0.25	0.22
29[ab]	-5.17	125.8	106	0.01	0.02	-4.07	127.9	110	0.13	0.02
30[ab]	-7.55	198.3	138	0.00	0.00	-7.53	193.8	142	0.02	0.00
31[ab]	-9.54	231.9	130	0.00	0.00	-9.46	260.8	147	0.00	0.00
32[ab]	-11.91	267.3	153	0.00	0.00	-12.61	298.5	142	0.00	0.00
33	0.00	181.2	167	0.02	0.04	0.00	161.0	153	0.00	0.29
34	-2.28	139.8	150	0.67	0.69	-2.62	162.1	165	0.06	0.37
35[b]	-3.72	128.5	109	0.03	0.06	-3.14	145.7	108	0.14	0.02
36	-2.67	167.7	159	0.15	0.21	-1.98	167.5	184	0.24	0.33

[a] $S\text{-}X^2$ statistics statistically significant for both ELL and non-ELL populations and using either the *HW* and *FDR* multiplicity corrections

[b] *Zh* test statistic statistically significant for both ELL and non-ELL populations (less than -2.998, the *HW* adjusted cut-off)

reported, both the pointwise and multiplicity adjusted values are provided. With regard to the multiplicity adjustments, Hunter-Worsley multiplicity adjusted values are reported (Hunter, 1976; Worsley, 1982) and the false discovery rate procedure (FDR) (Benjamini & Hochberg, 1995) used in order to provide conservative and liberal approaches to multiplicity adjustment.

Any differential fit for the ELL and non-ELL populations is assessed separately by testing for equal slope (loadings) of the IRT parameters. Checking for differential fit addresses another aspect of item equivalence of the CLASS III (3a). Only those items found to have an adequate level of fit based on the *Zh* and $S\text{-}X^2$ criteria and also not exhibiting differential fit between the ELL and non-ELL populations will be considered for the pool potentially having item and scalar equivalence. Note also that given the presence of the latent trait, the multidimensional polytomous IRT model assumes local independence. That is, conditional on the latent factors, the item responses should be independent. This assumption is checked in the analysis. Following assessment of the model fit, the individual items are evaluated for item (3a) and scalar (3b) equivalency by examining the discrimination parameters associated with each item and for each population (ELL and non-ELL). Finally, the scalar equivalence (3b) is assessed by examining the item characteristic curves (ICCs) of the individual items already found to be item equivalent. The ICCs summarize the likelihood of respondents selecting a particular level (e.g., 1 to 6) across all levels of the latent trait. Because the data are polytomous, then there is a curve for each level of the response. Whichever level (1 to 6) peaks at a particular point across the latent trait indicates more respondents chose that level. A well discriminating item will have all levels of the response represented (or peaking) across the latent trait in order.

Analysis of the CLASS instrument The proposed multidimensional polytomous IRT models are utilized to analyze the response patterns of the CLASS III items corresponding to the subscales identified in Section 2.1 for the CLASS III. These subscales include the dimensions of register (field, mode, and tenor) and any subscales of these dimensions as well (field: word confusion, technical words, and everyday words; mode: context of words, negative wording). In addition, all items included also met the following criteria: (1) applied to both ELLs and non-ELLs (e.g., CLASS III item 46: "If I learn a statistics concept in Spanish, I can easily work with it in English"), and (2) did not involve separate self-contained items (e.g., the textbook items did not qualify). For the scales examined, only the items that were on the common (i.e., for both ELLs and non-ELLs) part of the CLASS instrument were used.

Using items identified in Levels One and Two (see Appendix C) that show evidence for concept and construct operationalization equivalence, the CLASS items are investigated further to provide evidence of the item integrity. Recall that the identified subset of items presented fully in Wagler and Lesser (2014) align to a six-dimensional ordinal IRT model consistent with the three components of register and showing evidence of reasonable reliability. In order to investigate this set of promising items further, we utilized multiple group multidimensional IRT models to find evidence for or against item fit to the theoretical model.

Using all of these criteria in conjunction, it appears that eight items show strong evidence of misfit for both the non-ELL and ELL populations when assessing both fit measures (*Zh*, $S\text{-}X^2$). These are the items marked with an 'ab' superscript next to the item number in Table 4. Six items (14, 16, 18, 24, 26, 28, and 35) show evidence of misfit for both the non-ELL and ELL populations using the *Zh* criterion (which can be interpreted as a standard score and is compared to a Hunter-Worsley (*HW*) multiplicity corrected cutoff). These items are indicated by only having the 'b' superscript next to the item number. Only items 4 and 6 show evidence of misfit using the $S\text{-}X^2$ criterion alone as indicated by having only an 'a' superscript. The items with no superscripts appear to fit the theoretical model relatively well. The items showing strong evidence of misfit are discussed in detail in the following subsection where they are assessed for either deletion or revision in the CLASS IV.

In addition to examining the misfit statistics for the set of items, we assessed the discrimination parameters estimates. Appendix D presents the model thresholds and multivariate discrimination estimates, denoted *md,* in the table (Reckase, 2009). These multivariate discrimination estimates may be interpreted similarly as a discrimination estimate from a univariate IRT. Discrimination parameter estimates (*md*) outside of the bounds 0.5 and 2.0 are determined to be relatively weak discriminators. If *md* is less than 0.5, it is probably not discriminating enough and if it is greater than 2.0, then the

item has a highly bifurcating slope and merits further examination. By this criterion, item 4 shows low discriminatory power for non-ELLs but adequate for ELLs. Taking into account the feedback obtained from the focus groups, this item was dropped from further analysis and will not appear in the CLASS IV. No items had discrimination parameters greater than 2.0 for both ELL and non-ELL populations and, hence, all other items show evidence of adequate discriminatory power for both ELL and non-ELL populations. Note that item 28 does have a slightly high discrimination estimate (2.51) for the ELL but not for the non-ELL population.

Other items merit revision or deletion based on other criterion. For example, item 6 is already targeted for revision given the focus group feedback but may stay in the scale in revised form, and items 25 and 35 were selected for deletion by the focus groups. Following the assessment of these items showing evidence of misfit (and additionally not supporting the notion of item equivalence), we further investigate the subset of items determined to exhibit item equivalence. These are available upon request of the first author. The scalar equivalence of these items is investigated by assessing the degree of difference between the item characteristic curves (ICC) for each item. The ICCs for the six response levels are plotted for both the ELL and non-ELL populations. Using these ICC plots we can see how the latent trait of how the student perceives the role of language in statistics underlies their responses in the scale. Given the evidence of separate dimensions due to the field, tenor, and mode dimensions of register, these items are presented here by these categorizations for conceptual clarity. After excluding the items showing significant differences in the scalar properties between the ELL and non-ELL populations, the remaining items are still in consideration for inclusion in the CLASS IV. Table 5 summarizes the results of the qualitative and empirical analysis of the CLASS III with a focus on items still being considered for inclusion in the CLASS IV.

Table 5. Items with lack of cultural equivalence (Table 3) or statistical fit (Zh and S-X^2, Table 4), improper discrimination parameters (Appendix D), or discrepancy indicated between ICCs for ELL and non-ELL populations

Item	Table 3 data	*Zh* Misfit	*S-X*^{2} Misfit	Discrim-ination	ICC
4*			X	X	
6	X		X		
11*					
13*					
14	X	X			
15*		X	X		
16*		X			
18*		X			
19*					
21		X	X		X
22		X	X		X
24*		X			
26	X	X			
27		X	X		X
28*		X		X	
29*		X	X		
30		X	X		X
31		X	X		X
32		X	X		X
33*					
34*					
35		X			
36*	X				

*Indicates item was selected for inclusion in CLASS IV

The qualitative and quantitative evidence collected suggest that items marked with an asterisk in Tables 5 and 6 are conceptually and empirically stronger items than the others, and should be considered for further inclusion in the CLASS IV. After reviewing the evidence collected, the research team agreed upon the items. The items in Tables 5 and 6 that will no longer be included in the CLASS IV were eliminated for the following reasons. Item 6 was eliminated because the intent overlapped with items 4 and 33 and the item had a low estimated discrimination parameter in comparison to the other two items. Items 21, 22, 26, 27, 28, 30, 31, and 32 were eliminated because they all contained examples that unduly led students to rely upon the illustration without thinking generally about the concept. For example, in item 21, the phrase "e.g., the median and mean" will lead students to think only of that particular example of word confusion without considering the issue of word confusion with more generality. In this sense, items 21, 22, 26, 27, 28, 30, 31, and 32 were also not general enough for inclusion in the CLASS IV. In addition to these problems, the ICC curves associated with items 21, 22, 26, 27, 28, 30, 31, and 32 showed major evidence of discrepancies between the ELL and non-ELL populations. Taken altogether, these reasons justify exclusion from the CLASS IV. Finally, item 35 was eliminated due to ambiguous meaning and redundancy with other items, such as 4 and 11. Though item 14 shows evidence of misfit to the latent construct (via *Zh* misfit statistic) and the focus groups found the item problematic because they perceived students would be unwilling to criticize their professor, the research team felt this item was very important to include in the CLASS IV and have revised the wording to address these issues. Finally, item 24 was not included in CLASS IV because the referent of the word 'symbol' was not clear.

Table 6. Items included in the CLASS IV with description and revisions made

Item # from CLASS III	Register Dimension	What items assess	Revisions
4, 15, 16, 29*, 33*, 34	Field	How real-world context affects learning or vocabulary difficulties for statistical or non-statistical words	Common phrase from items 29 and 33 "Knowing the real-world situation" changed to "Being familiar with the real-world situation"
11, 13*, 19	Mode	Use of alternative modes of expressing statistical ideas-pictures, gestures, objects, news examples.	To clarify meaning, item 13 revised to: "Visual representations of statistical ideas are helpful."
14*, 18, 36	Tenor	How interaction between the student and professor may be affected due to linguistic/cultural background.	Lead wording on item 14 changed to "I wish professors waited more time…"
41*[a], 45*[a], 46*[a], 48*[a], 49*[a]	Field	How students transfer concepts between languages for statistics and everyday registers	All items were reworded to say "another language" when CLASS III used "Spanish"

[a]ELL-only item
*Indicates item is revised and included in the CLASS IV

The item order of the CLASS IV was chosen so that the first half of items had approximately equal characteristics as the second half of items, as recommended by DeVellis (1991). The characteristics of interest include the dimension of register (field, mode, or tenor) as well as statement length. These results support including three field items and either one or two mode and tenor items in each half of the set of items administered to both ELLs and non-ELL (because there are six field, three mode, and three tenor items). In the ELL-only set of items, the order was simply randomized because there are only five items total.

3.3. VALIDATION OF CLASS IV

The CLASS IV set of items was administered to 235 students enrolled in introductory statistics courses during spring 2015 and summer 2015. Out of the sample, each student was classified as being ELL or non-ELL using the Interagency Language Roundtable and American Council on the Teaching of Foreign Languages scale, where a 9 or 10 indicates a high level of English language proficiency and any score less than 9 indicates a lower level of English language proficiency. Using this scale, 93 students were classified as ELLs and 142 students were classified as non-ELLs. These students reflect the demographics in the original sample which was 40.5% ELL and reflects the widely reported statistic that approximately 40% of the student body has some level of limited English language proficiency.

The remainder of this survey consisted of the selected 12 questions as described in Section 3.4 for both ELL and non-ELL students, and 7 questions (also in Section 3.4) administered only to ELL students. It is necessary to validate the presumed structure of this scale for the population of interest. Using the sample data, we assess the structure of the data by conducting a modified parallel analysis (Timmerman & Lorenzo-Seva, 2011) to assess dimensionality and a polychoric factor analytic model to assess the loading structure. Both analyses make use of polychoric correlations as the data are ordinal (Olsson, 1979) and the factor analytic model will use the number of factors implied by the modified parallel analysis. The analysis is run on the combined pool of students and then separately run on ELL and non-ELL subsets in order to identify any scale non-invariancy with respect to language proficiency. A minimum residual factor rotation was applied in analyses presented. Additionally, all analyses were conducted in *R* (*R* Core Development Team, 2013) using the *psych* package (Revelle, 2015).

The results of the modified parallel analysis clearly indicate two dimensions to the updated set of CLASS items, hereafter called CLASS IV. Running a factor model with two factors yields fit statistics indicating a moderately good fit (χ^2=71.32 (*p*-value=0.0043), RMSEA interval (0, 0.043), RMSR = 0.04, and TLI = 0.956) and loadings indicate a clear and interpretable structure. However, the communalities ideally should be higher. Inspection of the polychoric correlation matrix does not indicate any redundancies among items (indicated by polychoric correlations greater than 0.80). Item 13 was deleted in order to investigate whether this item, which demonstrates a source of misfit, is inappropriately influencing the fit of the model. However, no dramatic changes are observed from the item loadings in Table 7 when item 13 is deleted. See Table 7 for a summary of the loading structure. Note that item numbers differ from those in the CLASS III.

Table 7. Items to be included in the CLASS IV with confirmatory factor analysis loadings and communalities

Item	Register Dimension	Factor 1	Factor 2	Communality
3. Visual representations of statistical ideas are helpful.	Mode: Mode of Words	0.32	-0.05	0.10
4. Statistical words in questions can make it difficult to answer.	Field: Technical Words	0.11	0.39	0.13
5. When a real-life situation illustrates the explanations of a concept, I feel there are now more opportunities for me to understand the concept.	Field: Everyday Context	0.47	-0.06	0.23
6. If I don't understand what is going on in class, I will pretend that I understand when the instructor looks towards me.	Tenor: Prof/Student Dynamics	0.07	0.42	0.18
7. In-class discussion of examples of statistics in the newspaper or a newsmagazine helps me understand statistics concepts.	Mode: Mode of Words	0.40	-0.17	0.19

8. Being familiar with the real-world situation the data comes from increases my understanding a sentence about statistical ideas.	Field: Everyday Context	0.63	0.02	0.40
9. When an instructor asks me a question in class, I believe that he/she thinks I know less than I really do because it takes me a while to express my thoughts into words.	Tenor: Prof/Student Dynamics	-0.15	0.39	0.17
10. It is confusing to me when words that look and sound similar (such as: mean, median, and mode) all get introduced in the same lesson.	Field: Word Confusion	-0.12	0.41	0.19
11. In class, I wish instructors waited more time so I can come up with an answer.	Tenor: Prof/Student Dynamics	0.00	0.58	0.33
12. Being familiar with the real-world situation the data comes from helps me understand the meaning of a statistical formula.	Field: Everyday Context	0.57	0.08	0.34
13. There are times I am not able to think of the correct academic words to describe something, but I am still able to communicate my understanding using gestures, pictures, or objects.	Mode: Modes of Words	0.36	0.25	0.19
14. Non-statistical words in questions can make it difficult to answer.	Field: Technical Words	0.05	0.38	0.15

Results from the construct analysis imply that item 13 should be revised and considered for deletion from the CLASS IV. Follow-up analysis with a focus group consisting of professors and graduate students in fields including biology, business, and education (held as a roundtable discussion at a 2016 international education conference) resulted in positive feedback about the scale content and potential use. The roundtable participants (one graduate student and two faculty members) attended the table because they themselves had been ELLs while taking university level introductory statistics. Two spoke Spanish as their first language and one spoke Hindi. At the roundtable, participants discussed questions regarding particular words in statistics and mathematics, including skew for example. Following discussion of particular terms, the attendees were asked to read through the CLASS IV items and provide feedback. One former Spanish-speaking ELL appreciated the framework of register theory and found the questions connecting expression of ideas using alternative modes for communication to be very relevant. The biologist attending the session noted that asking about real-world applications is important and relevant to her experiences in learning statistics. All agreed that they approved of the wording and content of the items.

4. DISCUSSION

4.1. SUMMARY

Understanding and respecting the cultural and linguistic diversity of our students is essential to providing quality instruction in statistics. As much of statistics instruction is in English and a substantial proportion of students do not speak English as their first language, this is an important consideration when designing and implementing instructional materials. We also note that many of the practices found using the CLASS (e.g., Lesser et al., 2013) are helpful for ELLs but also have positive impact for all students in introductory level statistics.

4.2. USE OF THE CLASS IV

The use of the CLASS IV is varied. Classroom instructors may observe a change in student demographics and want to better address the learning preferences, needs, and expectations of their class. In this case, the CLASS IV could be used to assist the individual instructor to guide adaptations in instruction. In addition, the CLASS IV could be used for research purposes. Some have explored how learning styles change based on cultural background (Verhoeven & Tempelaar, 2014; Mvududu, 2003), but few have investigated how classroom practices might change due to these factors or how language affects these differences. The CLASS IV could also be utilized in professional development settings to make statistics instructors aware of and better equipped to work with a changing student demographic. The scale is streamlined enough to be used in 5–10 minutes and the results can be readily interpreted by classroom instructors as well as by statistics education researchers.

4.3. LIMITATIONS AND DELIMITATIONS

The research study took place in the United States with a predominantly Hispanic bilingual population. This is a delimitation because the empirical results do not generalize beyond this population. However, establishing cross-cultural validity of the scale is an ultimate goal and qualitative results indicate it has validity for other student populations. Limitations of the study include the following. In the data collection, some respondents did not complete the survey and these results were not deemed "missing at random" and had to be eliminated. Other cases where the data appeared to be "missing at random" required imputation as detailed in the Methods section. Similarly, 42 respondents failed to indicate a language and these observations were also dropped and not analyzed. Finally, the focus groups ideally should have included participants from European and African countries, but no experts from these countries were available to be included in the focus groups.

4.4. DIRECTIONS FOR FUTURE RESEARCH

Future research should focus on administering the CLASS IV to diverse linguistic populations to identify how the structure may play out differently with students with linguistic backgrounds other than Spanish and English. The structure may vary depending on the cultural and linguistic background of the students despite the evidence presented in this paper about its cross-cultural relevance.

The results of the exploratory factor analysis presented in this paper may be used to inform regression analysis analyzing the relationship between the cultural and linguistic background and CLASS IV items. For each multiple regression, the response variable would be the mean of the items that significantly loaded for one of the five factors for ELLs identified in Table 3, and the independent predictor variables would be the largely non-numerical questions from the "student background" scale (i.e., items 1, 2, 3, 54, 55, 56a, 56b, 62). The goal is to identify which background variables are the most important predictors, in light of the observation that this ELL subpopulation is diverse in "length of residence in the US, language proficiency in English, language proficiency in Spanish, prior school experience, and socioeconomic status" (Moschkovich, 2003, p. 5). The CLASS instrument, however, did not ask about socioeconomic status and future uses of the scale should consider inclusion of this variable.

ACKNOWLEDGEMENTS

The authors express their appreciation to the focus group participants and to master's student Christie Mielke, who collected data and helped with data analysis for CLASS IV validation. The authors also appreciate the helpful editorial/reviewer feedback from *SERJ*.

REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the *R* environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chalmers, R. P., & Flora, D. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, *38*(5), 339–358.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86.

DeVellis, R. F. (1991). *Scale development: Theory and applications* (Vol. 26). London, UK: Sage Publications.

Dunn, P. K., Carey, M. D., Richardson, A. M., & McDonald, C. (2016). Learning the language of statistics: Challenges and teaching approaches. *Statistics Education Research Journal*, *15*(1), 8–27.
[Online: http://iase-web.org/documents/SERJ/SERJ15(1)_Dunn.pdf]

FSG Social Impact Advisors (2011). El Paso Regional Overview. 2017 Report available at: https://www.greatertexasfoundation.org/wp-content/uploads/2017/12/Research-TRAP-El-Paso-Full.pdf

Green, D. (1984). Talking of probability… *Bulletin of the Institute of Mathematics and its Applications*, *20*(9/10), 145–149.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Halliday, M. A. K., & Matthiessen, C. (2004). *Introducing functional grammar*. New York: Edward Arnold.

Hubbard, R. (1991). Teaching statistics to students who are learning in a foreign language. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics*, Dunedin, New Zealand (Vol. 1 pp. 514–517). Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/documents/papers/icots3/BOOK1/C10-6.pdf]

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology a review and comparison of strategies. *Journal of Cross-cultural Psychology*, *16*(2), 131–152.

Hunter, D. (1976). An upper bound for the probability of a union. *Journal of Applied Probability*, *13*(3), 597–603.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406.

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3), 1–19.
[Online: https://ww2.amstat.org/publications/jse/v17n3/kaplan.pdf]

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2010). Lexical ambiguity in statistics: How students use and define the words association, average, confidence, random and spread. *Journal of Statistics Education*, *18*(2), 1–22.
[Online: https://ww2.amstat.org/publications/jse/v18n2/kaplan.pdf]

Kaplan, J. J., Rogness, N. T., & Fisher, D. G. (2011). Lexical ambiguity: Making a case against spread. *Teaching Statistics*, *34*(2), 56–60.

Kaplan, J. J., Rogness, N. T., & Fisher, D. G. (2014). Exploiting lexical ambiguity to help students understand the meaning of random. *Statistics Education Research Journal*, *13*(1), 9–24.
[Online: http://iase-web.org/documents/SERJ/SERJ13(1)_Kaplan.pdf]

Kazima, M. (2006). Malawian students' meanings for probability vocabulary. *Educational Studies in Mathematics*, *64*(2), 169–189.

Lavy, I., & Mashiach-Eizenberg, M. (2009). The interplay between spoken language and informal definitions of statistical concepts. *Journal of Statistics Education*, *17*(1), 1–9.
[Online: https://ww2.amstat.org/publications/JSE/v17n1/lavy.pdf]

Lesser, L. (2010). Equity and the increasingly diverse tertiary student population: Challenges and opportunities in statistics education. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8),* Ljubljana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/documents/papers/icots8/ICOTS8_3G3_LESSER.pdf]

Lesser, L., Wagler, A., Esquinca, A., & Valenzuela, M. G. (2013). Survey of native English speakers and Spanish-speaking English language learners in tertiary introductory statistics. *Statistics Education Research Journal*, *12*(2), 6–31.
[Online: http://iase-web.org/documents/SERJ/SERJ12(2)_Lesser.pdf]

Lesser, L., & Winsor, M. (2009). English language learners in introductory statistics: Lessons learned from an exploratory case study of two pre-service teachers. *Statistics Education Research Journal*, *8*(2), 5–32.
[Online: http://iase-web.org/documents/SERJ/SERJ8(2)_Lesser_Winsor.pdf]

Liu, M. (2014). The social interpretation of language and meaning. *Theory and Practice in Language Studies, 4*(6), 1238–1242.

Lyberg, L. E., Biemer, P., Collins, M., De Leeuw, E. D., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (2012). *Survey measurement and process quality* (Vol. 1). New York: John Wiley & Sons.

Marín, G., & Marín, B. V. O. (1991). *Research with Hispanic populations* [Applied social research methods series]. Newbury Park, CA: Sage Publications.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Moschkovich, J. N. (2003). Understanding the needs of Latino students in reform-oriented mathematics classrooms In W. G. Secada, L. Ortiz-Franco, & N. G. Hernandez (Eds.), *Changing the faces of mathematics: Perspectives on Latinos* (pp. 5–12). Reston, VA: National Council of Teachers of Mathematics.

Mvududu, N. (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education*, *11*(3).
[Online: https://www.tandfonline.com/doi/full/10.1080/10691898.2003.11910726]

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Payán, R. M., & Nettles, M. T. (2008). Current state of English-language learners in the US K-12 student population. In 2008 English Language Learner Symposium, Princeton, NJ.
[Online: www.ets.org/Media/Conferences_and_Events/pdf/ELLsymposium/ELL_factsheet.pdf]

Pepitone, A., & Triandis, H. C. (1987). On the universality of social psychological theories. *Journal of Cross-Cultural Psychology*, *18*(4), 471–498.

Phillip, L., & Wright, G. (1977). Cultural differences in viewing uncertainty and assessing probabilities. In H. Jungermann & G. De Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 507–519). Dordrecht: Reidel.

R Core Team (2013). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137.

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, IL, USA.
[Online: https://cran.r-project.org/web/packages/psych/index.html]

Richardson, A. M., Dunn, P. K., Carey, M. D., & McDonald, C. (2016). Ten simple rules for learning the language of statistics. In H. MacGilivray, M. A. Martin, & B. Phillips (Eds.), *Proceedings of the 39th Australian Conference on Teaching Statistics (OZCOTS): Big Data: Mining, Analysing, Teaching* (pp. 32-37). Canberra, Australia: Statistical Society of Australia Inc.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220.

Valenzuela, M. G. (2009). *Survey research on communication and language for English learner and native English speakers enrolled in a college course on statistics literacy*. Unpublished master's thesis. The University of Texas at El Paso.

Verhoeven, P. S., & Tempelaar, D.T. (2014). Cultural diversity in statistics education: Bridging uniqueness. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8G3_VERHOEVEN.pdf]

Wagler, A., & Lesser, L. (2014). Assessing dimensionality of the Communication, Language And Statistics Survey: A multi-group analysis with introductory statistics students near the US-Mexico border. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/9/proceedings/pdfs/ICOTS9_C273_WAGLER.pdf]

Whitaker, D. (2016). Lexical ambiguities in the vocabulary of statistics. *International Journal for Mathematics Teaching and Learning*, *17*(3), 1–37.

Whitaker, D., Jaccobe, T., & Foti, S. (2014). Investigation of AP statistics students' understanding of technical terminology with possible lexical ambiguities. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://icots.info/9/proceedings/pdfs/ICOTS9_P13_WhitakerJacobbeFoti.pdf]

Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika*, *69*(2), 297–302.

AMY E. WAGLER
Department of Mathematical Sciences
The University of Texas at El Paso
500 W. University Avenue
El Paso, TX 79968
USA

APPENDIX A: COMMUNICATION, LANGUAGE, AND STATISTICS SURVEY (CLASS III)

1. What year in school are you?
   a) freshman b) sophomore c) junior d) senior e) graduate student
2. What kind of pre-service teacher are you?
   a) elementary school b) middle school c) high school d) I am not a pre-service teacher
3. About what percent of the material in this introductory statistics course do you estimate you already know on the first day of class? a) 0% b) 20% c) 40% d) 60% e) 80% f) 100%

Note: for items 4-36 and 40-49, the CLASS survey instructs the respondent to rate each statement using a six-point scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree) while also offering the option to choose "I don't understand."

4. When a real-life situation illustrates the explanations of a concept, I feel there are now more opportunities for me to understand the concept.
5. Understanding the statistical meaning of a word is difficult for me if that same word happens to mean something different in everyday conversational language.
6. Connections to words used in everyday conversation are most helpful to me when I encounter them before I encounter the technical academic terms.
7. If a statistics instructor says a word or phrase that I don't know, I am likely to stop listening for a moment while I turn and ask a neighbor or consult an aid such as a statistics dictionary.
8. I will understand a multi-word phrase used in statistics as long as I know each of the individual words in that phrase.
9. It is hard to tell whether a student does not understand a concept at all or whether that student has understanding but is not able to show it because one or more technical or academic words in the question are not familiar.
10. In-class discussion of examples of statistics in the newspaper or a newsmagazine helps me understand statistics concepts.
11. When a real-life situation illustrates the explanations of a concept, I feel there are now more words or ideas to have to read and understand to be able to understand the concept.
12. Working in groups in class helps me understand statistics concepts.
13. Using graphic organizers or pictures to organize my thinking is useful to me in statistics.
14. Professors often do not wait enough time after asking a question for me to think about what the question means, and think of an answer.
15. There have been times when I understood the concept, but was not able to answer a test question because I did not recognize some of the statistical words in the question.
16. There have been times when I understood the concept, but was not able to answer a test question because I did not recognize some of the nonstatistical words in the question.
17. If I did not understand the wording on a statistics test question (and if the wording was not a direct part of what was being tested), I would go up and quietly ask the professor during the test.
18. When a professor asks me a question in class, I believe that he/she thinks I know less than I really do because it takes me a while to express my thoughts into words.
19. There are times I am not able to think of the correct academic words to describe something, but I am still able to communicate my understanding using gestures, pictures, or objects.
20. It would be helpful if statistics instructors included examples that connect to my cultural background.
21. It is helpful when teachers explicitly distinguish statistics terms from words that may be unrelated but that sound the same or almost the same (e.g., median and medium).
22. It is helpful when teachers make analogies or connections between statistics words and real-world objects, such as: "just as a median divides a road into two halves (with opposite directions of travel), a median divides a dataset into two halves."
23. It is important to have discussions about statistical concepts in class that go beyond vocabulary definitions.
24. It is helpful when a teacher or a textbook takes the time to state how a new word or symbol is supposed to be pronounced.

25. [In statistics, the null hypothesis is what we assume is true before we collect data.] If the teacher asks whether the null hypothesis in a criminal courtroom trial is "defendant is innocent" or "defendant is guilty," the answer will have nothing to do with culture.
26. The phrase "not all group averages are equal" is difficult for me to understand.
27. The phrase "we failed to reject the null hypothesis" is difficult for me to understand.
28. The phrase "find the probability that no playing cards are not spades" is difficult for me to understand.
29. Knowing the real-world situation the data comes from helps me understand the meaning of words in a sentence involving statistical concepts.
30. It is confusing to me that some statistics words have several related slightly different words such as random, randomized, and randomization.
31. It is confusing to me that some statistics words are pronounced in different ways depending on the context, such as emphasizing the first syllable of survey (SURvey) when it's a noun and the second syllable (surVEY) when it's a verb.
32. It is confusing to me that statistics words such as random are used in a very different way in everyday speech than they are in a statistics class.
33. Knowing the real-world situation the data comes from helps me understand the meaning of a statistical formula.
34. It is confusing to me when words that look and sound similar (such as: mean, median, and mode) all get introduced in the same lesson.
35. Understanding the meaning of a statistical result is easier if I know the real-world situation the data comes from.
36. If I don't understand what is going on in class, I will pretend that I understand when the instructor is looking towards me.

37. What is your mother tongue? a) English b) Spanish c) other:_____

Look at your answer to question #37:

If it was "a" (English) or "c", you're now FINISHED with this survey. Thank you for your time.
If it was "b" (Spanish), please CONTINUE and answer the remaining questions.

38. Using the following (0-10) scale (based on scales used by the Interagency Language Roundtable and American Council on the Teaching of Foreign Languages), CIRCLE the number that best describes your proficiency with the English language:

Level	Description
10	Able to speak like an educated native speaker
9	Able to speak with a great deal of fluency, grammatical accuracy, precision of vocabulary and idiomaticity
8	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal or informal conversations
7	Able to satisfy most work requirements and show some ability to communicate on concrete topics
6	Able to satisfy routine social demands and limited work requirements
5	Able to satisfy most survival needs and limited social demands
4	Able to satisfy most survival needs and some limited social demands
3	Able to satisfy most survival needs and minimum courtesy requirements
2	Able to satisfy immediate need with learned utterances
1	Able to operate in only a very limited capacity
0	Unable to function in the spoken language

39. For each grade-level, select the column that indicates the language you were taught in for that grade. If you did not experience that grade or don't remember the language emphasis of that grade, leave that row blank.

	Taught mostly (or entirely) in English	Taught roughly equally in English and Spanish	Taught mostly (or entirely) in Spanish
First year of Jardin (3-year-olds)			
Second year of Jardin (4-year-olds)			
Third year of Jardin (5-year-olds; Kindergarten)			
First year of Primaria (1st grade in Elementary School)			
Second year of Primaria (2nd grade in Elementary School)			
Third year of Primaria (3rd grade in Elementary School)			
Fourth year of Primaria (4th grade in Elementary School)			
Fifth year of Primaria (5th grade in Elementary School)			
Sixth year of Primaria (6th grade in Elementary School)			
First year of Secundaria (7th grade in Junior High)			
Second year of Secundaria (8th grade in Junior High)			
Third year of Secundaria (9th grade in Junior High)			
First year of Preparatoria (10th grade in High School)			
Second year of Preparatoria (11th grade in High School)			
Third year of Preparatoria (12th grade in High School)			
First year of Escuela Superior (College)			
Second year of Escuela Superior (College)			
Third year of Escuela Superior (College)			
Fourth year of Escuela Superior (College)			

40. In statistics class, when I am working in a group with students who can speak in Spanish and in English, I would prefer to talk mostly in Spanish.
41. When I work by myself on a statistics problem, I usually think mostly in Spanish.
42. When the teacher mentions a word in Spanish that relates to the word I'm trying to learn in English, I find this to be helpful.
43. If I had an English-Spanish handbook of statistics terms (that does not give definitions or examples, but simply shows what statistics words in English correspond to what statistics words in Spanish), I would use it.
44. Most of what I already know about probability or statistics was learned in Spanish.
45. If I learn a statistics concept in English, I can easily work with it in Spanish.
46. If I learn a statistics concept in Spanish, I can easily work with it in English.
47. When a professor asks a question, I often translate it into Spanish for myself, figure out my answer, and then translate my answer back into English.
48. When I take a statistics test, I believe it would make a big difference if I had access to a list of matching statistics terms in Spanish and English.
49. When I take a statistics test, I believe it would make a big difference if I had access to a general English-Spanish dictionary to translate the "everyday" words used.

APPENDIX B: BACKGROUND ON CROSS-CULTURAL EQUIVALENCY EVIDENCE

Messick (1989) describes five conditions of validity for scores of a proposed scale: content validity, internal structure validity, criterion validity, response process validity, and consequences of use. It should be noted, though, that a basic assumption of any scale is that the scores are measuring a commonly understood theoretical construct. Multicultural settings tend to complicate, rather than simplify, the process of showing validity. In Section 2.2, the issues involved in a multicultural setting are discussed that expand the Messick framework of validity.

However, in addition to these factors affecting validity, we first consider the following: Among multi-ethnic populations, does English language proficiency of the respondents substantively affect the structure of the CLASS item responses? There is some guidance pertaining to Spanish-speaking Hispanic populations, such as how Marín and Marín (1991) advocate culturally sensitive approaches for understanding ethnic communities. Part of achieving a culturally sensitive analysis of the linguistic and cultural factors that affect learning in statistics is to adopt an etic-emic approach to the research study. That is, the construct grounding the CLASS III will be discussed utilizing a universal (*etic*) approach, while the response practices, cultural, and linguistic factors will be analyzed using a regional (*emic*) approach. With the regional approach in mind, this manuscript will describe and analyze heterogeneity present in the Spanish-speaking ELL population of primary interest and also assess any potential difficulties posed by linguistic differences. This approach both respects the role of the reference population of the scale (Spanish-speaking ELLs) while also allowing for cross-cultural comparison between ELLs and non-ELLs (as detailed in Section 2.2).

Secondly, whereas the CLASS focuses on the interaction of language in the learning of statistics, perhaps just as important is the role of culture and its impact on CLASS scores. By culture, we mean "a social group with a shared language and set of norms, values, beliefs, expectations, and life experiences" (Lyberg et al., 2012, p. 87). These two additional issues concerning the validity of the CLASS are certainly related to the components of validity as proposed by Messick (1989), but are separate enough to warrant separate treatment.

Comparisons across populations, however, must be made in valid and unbiased ways (Hui & Tirandis, 1985). For example, a cross-cultural study may hypothesize that cultural or linguistic factors exist when learning statistics, but any claimed difference must be demonstrated to be due to these factors and shown not to be artifacts of non-salient measurement differences between the populations. In general, the requirements for valid cross-cultural comparison include: conceptual and functional equivalence, equivalence in operationalization of the construct(s), item equivalence, and scalar equivalence (Hui & Triandis, 1985). In general, when assessing cross-cultural equivalence of scales, many researchers advocate utilizing multiple strategies and we note that the multiple strategies approach assumes a continuum with universality at one end and specificity at the other end. The universality of register theory (as discussed in Section 2.1) provides the necessary framework so that specific differences may be assessed and described using the scores from the scale. We note that demonstrating cultural equivalence of a scale is a hierarchical process where each level of equivalence (denoted as levels one, two, and three) must be met before proceeding to establish the next level of equivalence. This is analogous to using a sieve with an increasingly finer mesh at each level of equivalence. Hence, this manuscript follows the following multiple-strategy approach:

Level One: the cultural and functional equivalence of the CLASS III is argued based on the strong theoretical foundations of the scale, and

Level Two: the construct operationalization is assessed using existing factor analysis evidence of measurement invariance, and

Level Three: the item equivalence (3a) and scalar equivalence (3b) of the CLASS III is assessed using detailed item response analysis and using statistics measuring the differences among the intra-correlation coefficients (ICCs) across the populations.

This multiple-strategy approach encompasses macro-level (Levels One and Two) and micro-level (Level Three) ways of assessing how the CLASS III operates on multicultural ELL and non-ELL populations.

APPENDIX C: LOADINGS FROM FACTOR ANALYSIS ON CLASS III ITEMS (ELL; NON-ELL)

Confirmatory factor analytic models were utilized to test the CLASS IV item response dependencies and confirm the fidelity of the responses to the proposed register theory framework. The resulting factors can be categorized into the theorized domains of register: Field, Mode and Tenor. Item loadings for ELL and non-ELL populations and the corresponding uniqueness appear in the table (ELL; non-ELL).

	Field			Mode		Tenor
Paraphrased Items (uniqueness)	Everyday context (.36; .46)	Word confusion (.34; .43)	Technical words (.30; .11)	Negative wording (.51; .59)	Modes of words (.49; .41)	Professor-student dynamics (.37; .39)
4. Real-life context difficulties (.73; .63)	.66; .51					
5. Connection to everyday (.58; .67)	.55; .41					
6. Include discussion of vocabulary (.77; .75)					.51; .32	
13. Use graphic organizers (.58; .79)					.50; .61	
14. Not enough wait time (.64; .68)						.61; .68
15. No answer due to lack of words (.58; .39)			.77; .71			
16. No answer due to confusion about words (.15; .45)			.76; .84			
18. Professor thinks I know less due to words (.44; .57)						.63; .72
19. Student uses pictures (.83; .66)					.34; .39	
21. Confusion between registers (.44; .56)	.48; .64					
22. Real-life connection (.33; .41)	.73; .68					
23. Professor uses pictures (.43; .46)					.68; .65	
24. Help with pronunciation (.59; .63)					.57; .65	
26. Confusion about "not all means equal" (.21; .37)				.70; .91		
27. Confusion on "fail to reject H_0" (.40; .54)				.77; .70		
28. Confusion about "no playing cards are non-spades" (.48; .26)				.74; .71		
29. Real-world context difficult (.47; .67)	.45; .65					
30. Confusing similar words (.34; .45)		.69; .69				
31. Confusing pronunciations (.47; .39)		.66; .54				
32. Confusion about a specific word (.29; .22)		.98; .89				
34. Measures of center word confusion (.61; .64)		.34; .39				
35. Real-world context difficult (.42; .75)	.37; .72					
36. I pretend I understand (.63; .72)						.50; .59

Note. Under each factor, the proportion of variance appears (ELL; non-ELL)

APPENDIX D: DISCRIMINATION ESTIMATES FOR ELL AND NON-ELL POPULATIONS FROM THE MULTIVARIATE IRT MODEL

Using a multivariate item response model for more than two response levels, the item discrimination estimates are given in the table below. These correspond with each of the thresholds dividing the six response levels for each CLASS IV item. Also, the multidimensional discrimination (md) estimate is provided and can be interpreted in a standard manner.

	ELL						non-ELL					
Item	md	d1	d2	d3	d4	d5	md	d1	d2	d3	d4	d5
4	0.65	4.07	3.26	2.67	1.06	-0.89	0.47	4.03	2.85	2.05	0.91	-0.77
6	0.84	4.97	3.34	2.32	0.48	-1.86	0.65	5.09	3.28	2.17	0.74	-1.14
11	0.66	3.43	2.31	1.60	0.22	-1.74	1.06	3.81	2.51	1.81	0.45	-1.16
13	1.45	5.24	3.73	2.84	0.69	-1.23	1.21	5.03	3.64	3.17	1.45	-0.85
14	1.14	3.79	1.56	0.70	-0.89	-2.17	0.94	3.21	1.48	0.50	-0.57	-1.85
15	0.95	4.55	2.77	1.99	0.41	-1.39	1.16	4.44	2.81	2.23	0.64	-0.96
16	0.63	3.17	1.10	0.39	-0.88	-2.28	0.80	3.60	1.84	0.90	-0.13	-1.71
18	1.03	2.59	0.80	-0.05	-1.16	-3.17	1.26	2.60	1.19	0.10	-1.10	-2.92
19	0.93	3.20	1.72	1.10	-0.29	-2.20	0.61	3.74	2.11	1.46	0.07	-1.80
21	0.96	5.36	2.72	1.98	0.34	-1.53	1.20	4.49	3.35	2.76	1.38	-1.33
22	1.29	6.49	4.20	3.20	1.54	-1.02	1.53	6.70	4.66	3.63	2.25	-0.64
24	1.89	3.60	2.01	1.44	0.03	-1.96	1.44	4.71	3.19	1.92	0.37	-1.62
26	1.61	2.40	-0.03	-1.09	-2.37	-4.10	1.50	2.93	0.25	-0.78	-1.93	-3.85
27	1.68	2.85	0.38	-0.31	-2.12	-4.14	1.89	3.16	0.70	-0.28	-1.52	-3.11
28	2.51	3.03	0.54	-0.37	-1.73	-3.84	1.57	4.30	0.74	-0.10	-1.28	-2.84
29	0.99	5.63	3.70	2.67	0.72	-1.73	0.88	5.30	3.96	2.78	1.12	-1.43
30	0.98	3.51	1.35	0.55	-1.43	-3.10	0.90	4.04	1.68	0.91	-0.78	-2.60
31	0.86	2.15	0.26	-0.58	-2.15	-3.95	1.06	4.00	1.08	-0.07	-1.34	-2.97
32	1.17	3.31	0.36	-0.86	-2.77	-4.65	1.70	4.72	1.60	-0.14	-1.87	-3.97
33	0.00	2.56	1.09	0.63	-0.27	-1.79	0.00	2.74	1.45	0.90	-0.01	-2.17
34	0.57	1.93	0.22	-0.42	-1.63	-3.32	0.54	2.59	0.72	-0.11	-1.09	-2.43
35	0.76	4.33	3.05	2.21	0.23	-1.99	0.96	4.48	3.29	2.50	0.91	-1.60
36	0.90	2.73	1.34	0.70	-0.56	-2.02	0.65	2.32	1.23	0.57	-0.37	-1.58