CALIBRATION OF SELF-EFFICACY FOR CONDUCTING A CHI-SQUARED TEST OF INDEPENDENCE

WHITNEY ALICIA ZIMMERMAN
The Pennsylvania State University
waz107@psu.edu

DEBORAH D. GOINS
The Pennsylvania State University
dzl147@psu.edu

ABSTRACT

Self-efficacy and knowledge, both concerning the chi-squared test of independence, were examined in education graduate students. Participants rated statements concerning self-efficacy and completed a related knowledge assessment. After completing a demographic survey, participants completed the self-efficacy and knowledge scales a second time. Individuals with and without prior experiences with the topic were compared; those with prior experiences gave significantly higher self-efficacy ratings and had higher demonstrated knowledge scores, although the latter difference was not statistically significant. While self-efficacy and knowledge scores did not differ significantly between the two administrations, individuals without prior topic experience saw greater improvements in self-efficacy calibration. Findings suggest that self-efficacy calibration may be improved through completing an assessment.

Keywords: Statistics education research; Self-efficacy calibration; Statistics self-efficacy; Assessment

1. INTRODUCTION

The present study explored the relationship between self-efficacy and knowledge for students with and without prior experiences with a specific statistics topic: the chi-squared test of independence. According to Bandura (1977a, 1977b, 1978, 1997, 2007), *self-efficacy* is an individual's perceptions of his or her abilities to execute a given task to a certain level of success. "Perceived self-efficacy is concerned not with the number of skills you have, but with what you believe you can do with what you have under a variety of circumstances" (Bandura, 1997, p. 37). It is a task-specific measure of what an individual perceives him or herself to be capable of (Bandura, 2012). The task-specific nature of self-efficacy distinguishes it from other, broader constructs such as self-concept or self-esteem (Bandura, 1997; Pajares, 1996); self-efficacy measures have been shown to lead to better predictions of performance than more broad measures such as self-concept (Pajares & Miller, 1994).

There are four often cited influences on self-efficacy: modeling/vicarious experiences, personal experiences, psycho-physical status, and feedback from others (Schunk, 2012). In the present study, the impact of personal experiences is of primary interest. Self-efficacy is influenced by personal experiences with success or failure – that is, individuals attempt a

task, they succeed, and therefore, they are more likely to believe that they can be successful again in the future (e.g., Locke, Frederick, Lee, & Bobko, 1984).

In educational research, the relationship between self-efficacy and demonstrated knowledge is known as *self-efficacy calibration*, or simply, *calibration*. Calibration is the relationship between individuals' perceptions of their abilities and their abilities demonstrated in an observable manner (Glenberg, Sanocki, Epstein, & Morris, 1987; see also Alexander, 2013; Bembenutty, 2009). In the present study, these constructs were operationalized as self-efficacy and demonstrated knowledge, both concerning the topic of the chi-squared test of independence.

Research in a variety of disciplines has provided evidence for a relationship between self-efficacy and behaviors (Bandura, 1982; Schunk, 2012; Zimmerman, 2000). In general, learners with higher levels of self-efficacy are more likely to choose to engage in related behaviors – for example, college students enrolling in related courses or pursuing a related career path (Hackett & Betz, 1989; Lent, Lopez, & Bieschkie, 1991). Self-efficacy impacts how individuals view a task in terms of whether or not they expect to be successful; thus, individuals with higher levels of efficacy are often more likely to persist when challenged (Lent, Brown, & Larkin, 1984; Schunk, 1981).

Although these variables of engaging in related behaviors and persistence are positively correlated with self-efficacy, there may be some negative effects for students who have high levels of self-efficacy but notably lower skill levels – that is, those whose self-efficacy perceptions are not well calibrated with their actual abilities. They may have unrealistically high expectations of their performance and these may lead them to underprepare. Glenberg, Sanocki, Epstein, and Morris (1987), for example, note: "In preparing for a test of learning, a rational strategy is to study until one believes that the material is learned. Studying for less time is risky; studying for more time may be wasteful" (p. 119). Thus, there are practical implications of calibration that statistics instructors should consider.

Self-efficacy is also of great interest in the area of statistics education because of its connections with intrinsic motivation, persistence, performance, and self-regulatory behaviors (Bandura, 1982; Schunk, 2012; Schunk & Usher, 2011; Zimmerman, 2000; Zimmerman, Bandura, & Martinez-Pons, 1992). According to Pajares (1996), "the beliefs that individuals hold about their abilities and about the outcome of their efforts powerfully influence the ways in which they will behave" (p. 543). Self-efficacy is a construct that should be examined in greater detail in post-secondary statistics education settings. More specifically, as in the present study, students' self-efficacy calibration should be examined. If students' perceptions of their abilities are congruent with their actual abilities they will be able to make more informed decisions concerning their learning; furthermore, they will have accurate expectations.

In order to stay consistent with Bandura's (1977a, 1977b, 1978, 1997, 2007, 2012) definition of self-efficacy – which requires task specificity – the present study selected a topic comprised of multiple related tasks. The chi-squared test of independence was chosen to be the topic of focus in this research because it is an introductory statistics topic that is not taught until later in a second-semester statistics course at the college in which this study occurred. Because participants were sampled from a variety of courses on educational measurement, introductory statistics, and intermediate statistics, many would have experience with this topic from their previous coursework while many others would not yet have experience. Additionally, participants without direct prior experience with the chi-squared test of independence would still be able to draw from some of their general knowledge about the domain of statistics.

1.1. CALIBRATION AND EXPERIENCE

Although section 1.2 examines research concerning self-efficacy in statistics education settings, there is not extensive research specifically on calibration of self-efficacy and experiences in the domain of statistics. However, calibration and experiences have been examined in other settings. For example, research in the field of teacher education has shown that experiences can impact the relationship between self-efficacy and observable behaviors. In one study, Hoy and Spero (2005) measured teaching self-efficacy in pre-service teachers at the beginning of their teacher education program, at the end of their student teaching, and after one year of working as a full-time teacher. The researchers observed an increase in self-efficacy from the beginning of the program to the end of student teaching, but a decrease in self-efficacy from the end of student teaching to the end of the first year of full-time teaching. This decrease in self-efficacy was observed during a time in which it is assumed that their skills increased and suggests that calibration may be influenced by level of experience.

Again in the area of teacher education, Zimmerman, Parker, and Knight (2011) studied current teachers and the impact of a professional development experience on teaching efficacy and observable teaching behaviors. They found that at the end of the semester, for teachers who participated in the professional development program, there was a negative relationship between teaching efficacy and observable teaching behaviors while for teachers who were not exposed to the professional development program, the relationship was positive. This was tested using linear regression techniques: with group (professional development or control) dummy coded, there was a significant interaction between group and teaching efficacy in predicting observable constructivist teaching behaviors that were in line with the professional development program. The learning experience led to a change in the direction of the relationship between self-efficacy and observable teaching behaviors.

1.2. STATISTICS SELF-EFFICACY

In the domain of statistics, self-efficacy has been studied in terms of self-efficacy for learning statistics and self-efficacy for performing statistics. Research concerning self-efficacy for statistics-related tasks has examined its relationship with demonstrated knowledge; the findings suggest that a positive relationship exists between statistics self-efficacy and demonstrated knowledge and that statistics self-efficacy increases with related educational experiences (e.g., Finney & Schraw, 2003; Hall & Vance, 2010; Lane, Hall, & Lane, 2004; Zare, Rastegar, & Hosseini, 2011). Statistics self-efficacy has been operationalized in a number of ways. Some researchers have focused on students' perceptions of their current abilities (Finney & Schraw, 2003) while others have examined self-efficacy for successfully completing a statistics course (Bartsch, Case, & Meerman, 2012; Lane et al., 2004), self-efficacy for using a specific instructional method (Hall & Vance, 2010), or self-efficacy for learning in the domain of statistics (Chiesi, Primi, & Carmona, 2011; Finney & Schraw, 2003; Zare et al., 2011).

Finney and Schraw (2003), for example, measured statistics self-efficacy at the beginning and end of an introductory statistics course. They developed two parallel instruments for measuring students' self-efficacy for learning statistics and current statistics self-efficacy. Items included topics covered across an introductory statistics course, for example, "Identify the scale of measurement for a variable," and "Explain what the numeric value of the standard error is measuring" (p. 183), with each statement rated on a six-point scale (1 = *no confidence*, 6 = *complete confidence*). Demonstrated statistics knowledge was measured using course grades and scores on performance tasks that

consisted of items related to each of the items on the self-efficacy scales. Course grades and performance task scores both had the highest correlations with current statistics self-efficacy measured at the end of the course ($r = 0.496$ and 0.441, respectively). Self-efficacy for learning statistics at the beginning of the course was significantly positively correlated with course grades and performance task scores ($r = 0.340$, $r = 0.229$, respectively). Thus, the correlations between current self-efficacy and knowledge that were both measured at the end of the course were stronger than the correlations between self-efficacy for learning at the beginning of the course and measures of demonstrated knowledge at the end of the course. Changes in self-efficacy from the first to second administration were also examined; students' statistics self-efficacy increased on average by approximately two standard deviations ($t\ (109) = 18.64$, $p < 0.001$).

Lane et al. (2004) also examined statistics self-efficacy over time though over a shorter time period of approximately seven weeks. Like Finney and Schraw (2003), Lane et al. examined the relationships between self-efficacy and performance at various points during a statistics course. Their operationalization of self-efficacy was relatively broad; their instrument, the Self-Efficacy Toward Statistics Questionnaire (STSQ; see Lane, Hall, & Lane, 2002), included 44 items concerning six aspects related to success in a statistics course: statistical theory, lecture behaviors, information technology use, motivation, time management, and general competencies. Each item was rated on a five-point scale ($0 = not$ *at all confident*, $4 = very\ confident$). The six scales were summed to compute one overall STSQ score. STSQ scores did not change from the beginning to end of the study [$t(57) = 0.42$, $p = 0.34$]. Neither Lane et al.'s nor Finney and Schraw's research explicitly examined self-efficacy calibration.

Although the previously discussed studies examined statistics self-efficacy in more general terms, the present study took a topic-specific approach by examining self-efficacy for completing tasks related to the chi-squared test of independence. This is better in line with the task-specific nature of self-efficacy as a social cognitive construct as defined by Bandura (1977a, 1977b, 1978, 1997, 2007). It also allows for a more precise alignment between self-efficacy and demonstrated knowledge measures. Additionally, unlike previous research, the present study examines the relationship between self-efficacy and demonstrated knowledge with an emphasis on self-efficacy calibration. The possible impacts of different experiences (i.e., prior exposure to the topic and the presentation of a knowledge assessment) are also examined.

1.3. HYPOTHESES

The first research hypothesis is that self-efficacy ratings at the first and second administrations will differ for individuals with and without prior experiences with the topic of the chi-squared test of independence. The second research hypothesis is that the correlations between self-efficacy and demonstrated knowledge will differ at the first and second administrations for individuals with and without prior experiences with the topic of the chi-squared test of independence.

2. METHODS

2.1. PARTICIPANTS

Participants were recruited from graduate-level courses in introductory educational statistics, intermediate educational statistics, and intermediate educational measurement at a large research university in the northeast United States. Students who participated were given extra credit by their course instructors. Individuals who were unable to schedule a data collection session or who did not want to participate in the study were given the option of completing an alternative assignment for an equal amount of extra credit.

A total of 30 students participated. All participants were enrolled in masters or doctoral programs in the College of Education. The majority of the sample identified as female ($n = 22$) and the rest ($n = 8$) identified as male. Participants reported that they had completed between zero and five statistics courses ($\bar{X} = 1.933, SD = 1.484$). Half ($n = 15$) reported that they had previously learned about chi-squared tests in a statistics course.

2.2. INSTRUMENTS

Researcher-designed measures of demonstrated topic knowledge and self-efficacy, both on the topic of the chi-squared test of independence, were used in this study. Two sets (A and B) of the demonstrated topic knowledge were structured and designed to be approximately parallel. Set A of these items is presented in Appendices A and B. Varying item sets were used to prevent test-retest effects. Each form consisted of nine items that isolated the primary steps involved in conducting a chi-squared test of independence. Two forms of the demonstrated knowledge were created: the blue form presented set A followed by set B, and the red form presented set B followed by set A. This was done because it was not known if the two sets were parallel to one another. Participants were given a chi-squared table as well as the formulas for expected values and the chi-squared test statistic.

For each knowledge set, the first five items were presented in an online survey (see Appendix A). This format allowed participants to type their responses to these items which involved no or very minimal calculations. The last four items required more in-depth calculations (see Appendix B). Those four items were presented to participants in a paper packet.

All demonstrated knowledge items were open-ended. Two raters scored each response as 0 (for responses that were blank or completely incorrect), 1 (partially correct) or 2 (correct). After initial scoring, the two raters agreed on the scores to be assigned to 499 out of 540 responses (92.4%). The first author examined all scoring conflicts and made any necessary changes to be consistent with the agreed upon scoring rubric.

The self-efficacy scale consisted of nine items, each matching an item type on the demonstrated knowledge scale. For example, an item on the demonstrated knowledge scale asked participants to compute the test statistic for a 3x4 chi-squared test of independence. The matching item on the self-efficacy scale asked participants to rate their confidence in their current ability to compute the test statistic for a 3x4 chi-squared test of independence. This was done to establish consistency between the self-efficacy and demonstrated knowledge instruments.

At the beginning of the self-efficacy scale, participants were given the instructions, "Rate your confidence in your current ability to perform the following tasks related to **performing statistics** successfully." Each item was rated on a nine-point metric to be consistent with an additional scale that will be used in future research studies. The lower

end of the scale had the anchor "No Confidence." The upper end of the scale had the anchor "Complete Confidence." Participants were given the following instructions: "A rating of 1 is the lowest; give a rating of 1 to any tasks that you have no confidence that you could complete successfully. A rating of 9 is the highest; give a rating of 9 to any tasks that you have complete confidence that you could complete successfully." These anchors were derived from the scales used by Bartsch, Case, and Meerman (2012) and by Finney and Schraw (2003). The full scale as it appeared in the online survey is shown in Appendix C.

2.3. PROCEDURES

Data collection sessions were held in reserved computer labs in groups of six to ten participants at a time. Prior to the session, participants were sent the link to a password-protected survey. Participants were randomly assigned to receive the link to either the blue form of the survey or the red form of the survey. The only difference between the two forms was the order of presentation of the demonstrated knowledge item sets, as in Figure 1.
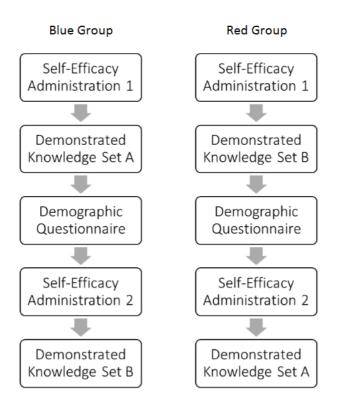

Figure 1. Procedures by group.

At the beginning of the data collection session, participants were given a password to access the online survey. After agreeing to the informed consent form presented at the beginning of the survey, the participants worked through each section of the survey as a cohort; that is, they started each section of the study at the same time as a group. First, participants completed items concerning their self-efficacy. Because each cohort of participants completed each section of the study at the same time, 75% completion was selected as a cutoff to ensure that the majority of participants completed each scale while not taking an excessive amount of time waiting for a small number of participants to

complete any section. After at least 75% of participants had submitted their responses, the researcher directed them to continue to the next section, which consisted of the first set of demonstrated knowledge items. Again, after at least 75% of participants had submitted their responses, the researcher directed them to continue to the next section of demographic questions. The demographic questionnaire was presented in the middle of the study as a brief buffer to separate the first and second administrations of the self-efficacy and knowledge scales. The same procedure was repeated as participants completed the self-efficacy items a second time and the last set of demonstrated knowledge items. Each data collection session lasted approximately one hour.

3. RESULTS

Results are presented in three sections. First, univariate statistics for the demonstrated knowledge and self-efficacy item responses are given for all participants combined; the reliability of scores from each scale is also examined. Second, changes in self-efficacy from the first to second administrations of the self-efficacy items are examined, for participants with and without prior topic-related experiences, in response to the first research question. Finally, the relationships between self-efficacy and demonstrated knowledge are examined, for participants with and without prior experience with the chi-squared test of independence, in response to the second research question.

3.1. UNIVARIATE STATISTICS

In this section, all participants, regardless of prior experiences, are included. Descriptive statistics for the demonstrated knowledge items for both set A and set B are presented in Table 1. Each item was rated on a scale of 0, 1, 2. Cohen's d was used as a measure of effect size to compare the difference in individuals' scores on set A versus set B. Positive effect sizes indicate that participants tended to score higher on set A; negative effect sizes indicate that participants tended to score higher on set B. For example, on the first item concerning identifying assumptions, on average participants scored 0.386 standard deviation higher on set A compared to set B. Overall, the mean score on set A was 7.633 ($SD = 3.358$, $n = 30$); the mean score on set B was 7.667 ($SD = 3.457$, $n = 30$). The effect size for the paired difference between set A and set B was $d = 0.020$.

Table 1. Demonstrated knowledge item descriptive statistics

Item Type	Set A		Set B		Cohen's d
	Mean	SD	Mean	SD	
Identify assumptions	0.300	0.466	0.167	0.379	0.386
Write null and alternative hypotheses	0.133	0.507	0.333	0.758	-0.248
Look up critical value	1.267	0.980	1.267	0.980	0
Calculate degrees of freedom	0.133	0.507	0.133	0.507	0
Given a χ^2 test statistic, decide whether to reject the null hypothesis	0.967	0.928	1.200	0.887	-0.286
Calculate row and column totals	1.900	0.403	1.967	0.183	-0.148
Calculate expected cell values	1.168	0.986	1.100	0.995	0.077
Compute the test statistic for 2x2 χ^2 test of independence	0.967	0.890	0.867	0.937	0.151
Compute the test statistic for a 3x4 χ^2 test of independence	0.800	0.761	0.633	0.765	0.281

Additionally, to assess the parallel forms of the knowledge measure, a repeated measures analysis of variance was used to compare scores on the two item sets for participants who received them in varying orders. The interaction between form (blue and red) and item set (A and B) was not significant (Wilks' lambda = 0.996, F (1, 28) = 0.103, p = 0.751, η_P^2 = 0.004). There was not a significant main effect for item sets (Wilks' lambda = 1.000, F (1, 28) = 0.008, p = 0.931, $\eta_P^2 < 0.001$) nor was there a difference between the scores of participants who received the AB (blue) form versus the BA (red) form (F (1, 28) = 0.154, p = 0.698, η_P^2 = 0.005). Because the order in which set A and set B items were presented to participants varied by form, all further analyses use composite scores which combine all items from item sets A and B combined. The resulting 18-item scale yielded a Cronbach's alpha of .810. The correlation between composite scores on set A and set B was r = 0.883 ($p < 0.001$, n = 30).

Descriptive statistics for each of the self-efficacy items at the first and second administration are presented in Table 2. Participants rated each item on a scale of 1 to 9. The effect sizes for the changes in ratings are also given in Table 2. Positive effect sizes indicate increases in self-efficacy ratings; negative effect sizes indicate decreases in self-efficacy ratings. For example, participants' ratings on the first item, "Identify the assumptions of a χ^2 test of independence," decreased on average by 0.533 standard deviation from the first to second administration. The correlation between composite self-efficacy scores at the first and second administrations was r = 0.699 ($p < 0.001$, n = 30). In terms of reliability of ratings, Cronbach's alpha was 0.954 at the first administration and 0.933 at the second administration.

Table 2. Self-efficacy item descriptive statistics

	Administration 1		Administration 2		*Cohen's d*
	Mean	SD	Mean	SD	
Identify the assumptions of a χ^2 test of independence	3.38	1.840	2.67	1.807	-0.533
Write the appropriate null and alternative hypotheses for a χ^2 test of independence	4.00	2.375	2.80	2.058	-0.602
Look up the appropriate critical value on a χ^2 table	4.59	2.666	5.00	2.729	0.213
Calculate the appropriate degrees of freedom for a given χ^2 test of independence	4.34	2.609	3.23	2.161	-0.488
Calculate row and column totals	5.21	2.808	6.73	2.586	0.562
Calculate expected cell values	3.93	2.604	4.73	2.912	0.359
Compute the test statistic for a 2x2 χ^2 test of independence	3.21	2.161	3.70	2.680	0.257
Compute the test statistic for a 3x4 χ^2 test of independence	3.03	2.129	3.60	2.724	0.265
Given a χ^2 test statistic, make the correct decision whether to reject or fail to reject the null hypothesis	4.86	2.887	4.57	3.014	-0.074

3.2. HYPOTHESIS 1

Recall that the first research hypothesis was: *The self-efficacy ratings given at the first and second administrations will differ for individuals with and without prior experiences with the topic.* The self-efficacy ratings of participants with and without prior experience

with the chi-squared test at the first and second administrations of the scale were compared using a repeated measures analysis of variance. Prior experience served as a between-subjects independent variable; administration served as a repeated within-subjects independent variable. Composite scores on the nine self-efficacy items served as the dependent variable. Descriptive statistics are presented in Table 3. There was not a statistically significant interaction between prior experience and administration (Wilks' lambda = 0.954, F (1, 28) = 1.345, p = 0.256, η_p^2 = .0046), nor was there a main effect for administration (Wilks' lambda = 0.991, F (1, 28) = 0.261, p = 0.614, η_p^2 = 0.009). Thus, there was not evidence that self-efficacy changed following the assessment experience. However, there was a significant main effect for prior experience (F (1, 28) = 17.890, p < 0.001, η_p^2 = 0.390): individuals with prior experience with the chi-squared test of independence gave significantly higher ratings on the self-efficacy items and the effect size was very large.

Table 3. Self-efficacy by administration and prior experience

	N	Administration 1		Administration 2	
		Mean	SD	Mean	SD
With prior experience	15	47.933	13.424	46.200	15.289
Without prior experience	15	23.408	16.502	27.967	17.250

3.3. HYPOTHESIS 2

Recall that the second research hypothesis was: *Correlations between self-efficacy and demonstrated knowledge will differ before and after an assessment experience for individuals with and without prior experiences with the topic.* As seen in Table 4, the correlation between demonstrated knowledge and self-efficacy increased from the first to second administration for individuals with and without prior experience with the topic. Changes in the relationship between demonstrated knowledge and self-efficacy from the first to second administrations were examined for both groups; a directional, dependent correlational comparison with one variable in common was used (Lee & Preacher, 2013). It was hypothesized that the correlations between demonstrated knowledge and self-efficacy would increase from the first to second administrations. For individuals with prior experience with the chi-squared test of independence, the increase in correlation between demonstrated knowledge and self-efficacy approached statistical significance (z = 1.63, p = .051). For individuals without prior experience with the topic, the increase in the correlation was statistically significant (z = 3.38, p < .001).

Table 4. Correlations by prior topic experience

		1	2	3
1.	Demonstrated Knowledge	--	0.149	0.441
2.	Self-Efficacy Administration 1	0.037	--	0.776*
3.	Self-Efficacy Administration 2	0.800*	0.418	--

Values above the main diagonal are for participants (n=15) with prior experience with the χ^2 test of independence; values below the diagonal are for participants (n=15) without prior experience.
* p < 0.01 (2-tailed)

4. DISCUSSION

4.1. SUMMARY OF FINDINGS

Overall, self-efficacy ratings at the first and second administration were not significantly different. Unlike previous studies where changes in statistics self-efficacy have been observed (e.g., Finney & Schraw, 2003), the present study involved no provided instruction and had data collected in one session as opposed to over multiple weeks. These results provide evidence for the stability of scores over a brief period of time. At both administrations, individuals with prior experience with the topic gave significantly higher ratings on the self-efficacy items compared to individuals without prior related experiences.

The most interesting finding may be the changes in the relationships between self-efficacy and knowledge, for individuals with and without prior experiences, from the first to second administrations. Both groups saw an increase in the correlations over the course of this short study. Individuals without related prior experiences saw a greater increase (r = 0.037 to 0.800) than individuals with related prior experiences (r = 0.149 to 0.441). When converted to proportions of explained variance (R^2), individuals without prior experience saw a greater increase ($\Delta R^2 = .06386$) than those with prior experience ($\Delta R^2 = 0.1723$) from the first to second administrations of the self-efficacy items. The previously reviewed statistics self-efficacy studies did not examine such changes in correlations (e.g., Finney & Schraw, 2003; Lane et al., 2004). These findings are consistent with the notion that self-efficacy is most impressionable in novices (e.g., Hoy & Spero, 2005).

4.2. LIMITATIONS

All participants in this study were graduate-level students enrolled in educational statistics or measurement courses. The relatively homogeneous sample does limit the generalizability of results to other populations, such as that of undergraduate-level students who have different educational backgrounds and possibly different motivation levels. Additionally, the sample was not random – students were recruited from existing courses and they volunteered to participate. It is not known if the sample of students who participated is representative of all students in those courses or at the college. Future studies are planned that will examine self-efficacy for completing tasks related to the chi-squared test of independence in more heterogeneous samples including undergraduate students from a variety of academic majors. A second limitation of this study was the relatively small sample size. With a total sample size of 30, and only 15 in each category, results should be interpreted with caution.

Furthermore, the variable of previous experience was not manipulated by the researchers in this study. Instead, participants were asked to self-report their prior experience. There may have been confounding variables; for example, participants with prior experience with the topic of the chi-squared test of independence probably had more experience in the domain of statistics overall.

Finally, another limitation was its focus on the procedural knowledge associated with conducting a chi-squared test of independence. Most tasks included on the self-efficacy and demonstrated knowledge measures emphasized computations. Many of these tasks require the use of mathematics and thus, mathematics ability may have been a confounding variable. Results may have been different if the instruments had more emphasis on conceptual knowledge.

4.3. SUGGESTIONS FOR FUTURE RESEARCH

Future studies should continue to explore self-efficacy for completing specific tasks, particularly those that are often challenging for statistics students. The impact of self-efficacy calibration on self-regulatory behaviors is of great interest because of its implications for practice. Interventions for improving self-efficacy calibration and self-regulatory behaviors should be examined with the goal of developing suggestions for practice. For example, the findings of the present study suggest that providing introductory statistics students with practice tests before an exam may help them to judge better their current abilities and therefore inform their study habits. Future research may examine how different types of practice assessments (e.g., self-evaluated, instructor-evaluated, etc.) impact students' perceptions of their abilities and how they impact their studying behaviors. Additionally, in the present study, participants were not given the correct answers nor were they informed of their level of performance. Future research should examine the impact of practice tests that include such feedback.

4.4. CONCLUSIONS

Because self-efficacy is explicitly a task-specific construct (Bandura, 1997), it was necessary to develop measures of self-efficacy and demonstrated knowledge that focused on individual tasks. This contrasts with previous research studies which have examined statistics self-efficacy on a broader basis (e.g., Bartsch et al., 2012; Finney & Schraw, 2003; Lane et al., 2004).

The results of this study suggest that an assessment experience may influence the relationship between self-efficacy and knowledge. Increases in the correlation between self-efficacy and demonstrated knowledge were observed, for both participants with and without prior experience with the topic, following the directly related assessment experience; this increase was greater for individuals without prior topic experience. Such a finding reveals the positive impact of a brief exercise (involving nine knowledge items). This may help instructors when determining ways to provide supports in order to maximize student learning behaviors. For example, providing practice tests may help students form more accurate interpretations of their actual abilities. Although many instructors already incorporate practice assessments into their courses for a variety of reasons (e.g., reducing anxiety, providing information about the format of an exam), the present study suggests that practice tests may also be related to self-efficacy calibration, which in turn may impact self-regulatory behaviors such as determining what to study and how much to study. These findings are promising, but more research on the impact of practice tests and self-efficacy calibration on self-regulatory behaviors is needed.

Whether it is an introductory undergraduate-level statistics course or a more advanced graduate-level statistics course, students bring varying levels of prior knowledge and experiences. Students' perceptions of their abilities may contribute to outward behavioral applications of skills and time invested in learning activities. The frequency and time period of opportunities to demonstrate learning may affect not only the observable outcomes but also students' perceptions of their abilities. Taking prior knowledge into account, a closer examination of the relationship between self-efficacy, self-regulatory behaviors, and demonstrable knowledge may ultimately enhance course design to maximize student learning. Therefore, a better understanding of how self-efficacy is related to previous experiences, assessment experiences, and demonstrated knowledge is important for understanding student learning.

ACKNOWLEDGMENT

REFERENCES

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction*, *24*, 1–3.

Bandura, A. (1977a). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review*, *84*(2), 191–215.

Bandura, A. (1977b). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bandura, A. (1978). Reflections on self-efficacy. In S. Rachman (Ed.), *Advances in Behavioral Research and Therapy* (Vol. 1, pp. 237–269). Oxford: Pergamon.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122–147.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman.

Bandura, A. (2007). Much ado over a faulty conception of perceived self-efficacy grounded in faulty experimentation. *Journal of Social and Clinical Psychology*, *26*(6), 641–658.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management*, *38*(1), 9–44.

Bartsch, R. A., Case, K. A., & Meerman, H. (2012). Increasing academic self-efficacy in statistics with live vicarious experience presentation. *Teaching of Psychology*, *39*(2), 133–136.

Bembenutty, H. (2009). Three essential components of college teaching: Achievement calibration, self-efficacy, and self-regulation. *College Student Journal*, *43*(2B), 562–570.

Chiesi, F., Primi, C., & Carmona, J. (2011). Measuring statistics anxiety: Cross-country validity of the Statistical Anxiety Scale (SAS). *Journal of Psychoeducational Assessment*, *29*(6), 559–569.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161–186.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*(2), 119–136.

Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, *20*(3), 261–273.

Hall, S., & Vance, E. A. (2010). Improving self-efficacy in statistics: Role of self-explanation & feedback. *Journal of Statistics Education*, *18*(3).
[Online: www.amstat.org/publications/jse/v18n3/hall.pdf]

Hoy, A. W., & Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and Teacher Education*, *21*(4), 343–356.

Lane, A. M., Hall, R., & Lane, J. (2002). Development of a measure of self-efficacy specific to statistics courses in sport. *Journal of Hospitality, Leisure, Sport and Tourism Education*, *1*(1), 47–56.

Lane, A. M., Hall, R., & Lane, J. (2004). Self-efficacy and statistics performance among sport studies students. *Teaching in Higher Education*, *9*(4), 435–461.

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common.

[Online: http://quantpsy.org/corrtest/corrtest2.htm]

Lent, R. W., Brown, S. D., & Larkin, K. C. (1984). Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology*, *31*(3), 356–362.

Lent, R. W., Lopez, F. G., & Bieschkie, K. J. (1991). Mathematics self-efficacy: Sources and relation to science-based career choice. *Journal of Counseling Psychology*, *38*(4), 424–430.

Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, *69*(2), 241–251.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*(2), 193–203.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology*, *73*(1), 93–105.

Schunk, D. H. (2012). Social cognitive theory. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, J. Sweller, & J. Brophy (Eds.), *APA educational psychology handbook* (pp. 101–124). Washington, DC: American Psychological Association.

Schunk, D. H., & Usher, E. L. (2011). Assessing self-efficacy for self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 282–297). New York: Routledge.

Zare, H., Rastegar, A., & Hosseini, S. M. D. (2011). The relation among achievement goals and academic achievement in statistics: The mediating role of statistics anxiety and statistics self-efficacy. *Procedia, Social and Behavioral Sciences*, *30*, 1166–1172.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82–91.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, *29*(3), 663–676.

Zimmerman, W. A., Parker, D., & Knight, S. L. (2011). *Professional development's effect on the relationship between efficacy and behaviors in reforming Qatar*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

WHITNEY ALICIA ZIMMERMAN
Department of Statistics
The Pennsylvania State University
325 Thomas Building
University Park, PA 16802

APPENDIX A. DEMONSTRATED KNOWLEDGE MEASURE PART I

1. List two assumptions of the χ^2 test of independence

1: _____

2: _____

2. What is the appropriate _null_ hypothesis for a 2x3 χ^2 test of independence?

H_0:

_____

3. You are conducting a χ^2 test of independence with 5 degrees of freedom and a .01 level of significance. What is your critical value?

_____

4. In a 5x6 χ^2 test of independence, what is the numerical value of the degrees of freedom?

_____

5. You conducted a χ^2 test of independence with two independent variables. The critical value from the χ^2 table is 5.99 and your χ^2 test statistic is 4.35. What is your conclusion concerning the independence of the two variables? Why?

_____

APPENDIX B. DEMONSTRATED KNOWLEDGE MEASURE PART II

6. Compute the row and column totals for the following table. (i.e., fill in the empty cells)

		Sex		Row Totals
		Males	Females	
Eye Color	Blue	20	40	
	Brown	30	50	
	Other	40	20	
	Column Totals			200

7.

	Choice A	Choice B
Outcome 1	10	20
Outcome 2	40	30

Using the data in the table above, compute the expected cell values and write them in the table below. Assume that choice and outcome are independent of one another.

	Choice A	Choice B
Outcome 1		
Outcome 2		

8. Compute the χ^2 test statistic for the following table. The expected values are in parentheses.

		Coin Flip		Row Totals
		Heads	Tails	
Coin Age	New	10 (8)	8 (10)	18
	Old	6 (8)	12 (10)	18
	Column Totals	16	20	36

9. Compute the χ^2 test statistic for the following table. The expected values are in parentheses.

	Option A	Option B	Option C	Option D	Row Totals
Left-Handed	8 (10)	11 (8)	12 (10)	9 (12)	40
Right-Handed	23 (25)	22 (20)	30 (25)	25 (30)	100
Ambidextrous	19 (15)	7 (12)	8 (15)	26 (18)	60
Column Totals	50	40	50	60	200

APPENDIX C. SELF-EFFICACY SCALE

Rate your confidence in your current ability to perform the following tasks related to **performing statistics** successfully.

A rating of 1 is the lowest; give a rating of 1 to any tasks that you have no confidence that you could complete successfully.
A rating of 9 is the highest; give a rating of 9 to any tasks that you have complete confidence that you could complete successfully.

	No Confidence							Complete Confidence	
	1	2	3	4	5	6	7	8	9
Identify the assumptions of a χ^2 test of independence	○	○	○	○	○	○	○	○	○
Write the appropriate null and alternative hypotheses for a χ^2 test of independence	○	○	○	○	○	○	○	○	○
Look up the appropriate critical value on a χ^2 table	○	○	○	○	○	○	○	○	○
Calculate the appropriate degrees of freedom for a given χ^2 test of independence	○	○	○	○	○	○	○	○	○
Calculate row and column totals	○	○	○	○	○	○	○	○	○
Calculate expected cell values	○	○	○	○	○	○	○	○	○
Compute the test statistic for a 2x2 χ^2 test of independence	○	○	○	○	○	○	○	○	○
Compute the test statistic for a 3x4 χ^2 test of independence	○	○	○	○	○	○	○	○	○
Given a χ^2 test statistic, make the correct decision whether to reject or fail to reject the null hypothesis	○	○	○	○	○	○	○	○	○