*Journal of the Scholarship of Teaching and Learning, Vol. 16, No. 4, August 2016, pp. 75-91.*
*doi: 10.14434/josotl.v16i4.19371*

# Using Indirect vs. Direct Measures in the Summative Assessment of Student Learning in Higher Education

**Christine Luce**[1] **and Jean P. Kirnan**[2]

*Abstract: Contradictory results have been reported regarding the accuracy of various methods used to assess student learning in higher education. The current study examined student learning outcomes across a multi-section and multi-instructor psychology research course with both indirect and direct assessments in a sample of 67 undergraduate students. The indirect method measured student perceived knowledge and abilities on course topics, while the direct method measured actual knowledge where students answered test questions or solved problems reflecting course content. Both measures independently demonstrated increases from pretest to posttest; however only the direct measure correlated with final course grades. Results also showed respondents scoring lower on the direct measure were overconfident (as measured by indirect score) in their perceived knowledge and ability, the Dunning-Kruger Effect. Based on our findings, we concluded that the indirect method was not an accurate measure of student learning, but may have benefits as an instructional tool.*

*Keywords: indirect and direct measures, summative assessment, pretest/posttest, psychology, perceived knowledge, confidence, student learning outcomes, higher education*

The support of student learning is a primary goal of higher education. Colleges attempt to monitor student learning by implementing assessments across multiple levels within their institutions (Banta, 2004). However, the nature of these assessments has shifted in recent years from assessing faculty teaching to demonstrating student learning (Martell, 2007). Increasingly, institutions of higher education are being challenged by their accrediting agencies (Martell, 2007) and legislative entities (Fort, 2011) to not only provide evidence of the investment of resources and offering of programs for various learning initiatives, but to also demonstrate actual student learning and the achievement of specific program learning outcomes. Such a challenge requires consideration of many factors, including when and where to assess student learning, how to configure meaningful samples, and, most critically, how to decide on the best measure of learning.

Many methods have been suggested for measuring student learning including standardized tests (locally developed or commercially available), evaluation of current student work, portfolio assessments, and self-report surveys by current students and alumni. The actual determination of the "best" method depends on several factors, such as program learning objectives, program size, course sequence, and institutional resources to support assessment. Departments considering the development of their own assessment tools often consider the advantages of direct measures vs. indirect measures of student learning. Direct measures of student learning demonstrate mastery through actual work or work products such as papers, presentations, embedded test items, and pretests and posttests. Indirect measures, on the other hand, reflect attitudes or opinions, and are often obtained from focus groups, interviews, or surveys (Price and Randall, 2008). This study focused on the measure of assessment, contrasting the use of direct methods and indirect methods

[1]Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541
[2]Department of Psychology, The College of New Jersey, 2000 Pennington Road, Ewing, NJ 08628

of determining student learning in a psychology research methods course at a higher education institution.

## Literature Review

### *Direct Measures*

Direct measures of learning demonstrate mastery of knowledge rather than state an opinion about one's abilities. There are numerous methods of direct measurement including objective tests, embedded test items, papers, projects, and presentations. These methods can constitute newly developed measures, purchased measures, or existing student work as long as the tasks involved reflect the learning outcomes targeted for assessment.

The use of direct measures administered at the beginning and at the end of a semester can determine course or program value-added by identifying increased knowledge from the start to the end of a course (Pederson and White, 2011) as well as pinpoint areas of weakness (Price and Randall, 2008). Pretest direct measures identify a student's prior knowledge and skills, and the results can also inform the need for course content modification to prevent overlap in course topics (Price and Randall, 2008).

Direct measures that are developed locally are time-consuming to create and administer. If developed by faculty teaching the course, local measures also run the risk of reflecting a specific pedagogy rather than broad learning objectives. Commercially available measures benefit from professional development, large sample sizes and the opportunity for peer institution comparison (Office of the Provost, 2015). Commercial measures can also be expensive, and may lack alignment with a specific institution's curriculum or learning outcomes.

Furthermore, several researchers have recently identified concerns regarding student motivation when taking standardized measures constructed for assessment purposes (Huffman, Adamapoulos, Merdock, Cole, and McDermid, 2011; Liu, Bridgeman, and Adler, 2012). The "high stakes" for an institution or department contrast with the "low stakes" for the student, and research has shown that a lack of motivation on the part of respondents may influence the validity of these measures as well as the conclusions drawn regarding student learning. Huffman et al. (2011) demonstrated higher scores on the Psychology Area Concentration Achievement Test (a major fields test) when students were provided with a statement on the importance to the department of honest and effortful responses. Liu et al. (2012) found similar results for the ETS Proficiency Profile where higher scores were achieved when an institutional or personal motivation condition was present. These researchers collectively resolved that assessment without proper respondent motivation can lead to erroneous conclusions.

The use of existing student work could overcome many of the motivational issues raised above as student assignments traditionally carry some "weight" for a final grade, and therefore constitute "high stakes" for the student, resulting in increased motivation. However, in the department under study, while course content was consistent, student tasks differed significantly across the multiple course sections and instructors. Thus, the use of existing student work was not deemed feasible due to its variability. Instead, we used a direct measure of student learning which was independently developed to reflect consistent course content and we administered the measure in a standardized manner to students in all course sections.

### *Indirect Measures*

Collecting data indirectly may involve focus groups, student interviews, instructor and course ratings, or knowledge surveys (Price and Randall, 2008). Much research has been conducted on the use of knowledge surveys which require students to indicate their perceived confidence or ability to answer questions, but do not require actual answers or problem solving (Nuhfur and Knipp, 2003). Due to the time savings from not having to actually solve problems, hundreds of items can be used to cover a wide array of course topics, an option not feasible with direct measures (Nuhfer and Knipp, 2003; Price and Randall, 2008).

Benefits from these types of knowledge surveys can be seen by both the faculty and the student (Nuhfer and Knipp, 2003). Details of course content reviewed prior to the semester aid in course preparation and expectations (Nufher and Knipp, 2003). When used in a pretest/posttest methodology, knowledge surveys can provide instructors with the ability to analyze course content and learning objectives, allowing them to modify instruction and ultimately improve student learning (Wirth and Perkins, 2005).

For students, the pretest survey can be used as a preview of the semester's course content (Nufher and Knipp, 2003) and further used as a study guide throughout the semester (Bell andVolckmann, 2011; Bowers, Brandon, and Hill, 2005; Clauss and Geedney, 2010; Nufher and Knipp, 2003; Wirth and Perkins, 2005). Indirect measures can also afford students the opportunity and ability to recognize and further develop their self-assessment skills (Bell and Volckmann, 2011; Clauss and Geedey, 2010; Wirth and Perkins, 2005).

There are, however, some drawbacks to the use of indirect measures such as knowledge surveys. Traditionally presented as Likert scale items, knowledge surveys may be susceptible to response sets such as social desirability, impression management, and "yea-saying," especially if students are concerned that faculty will have access to their responses. One could postulate that the issue of motivation noted earlier for the direct measures likely applies to indirect measures as well, even though these measures are less onerous for the respondent to complete. Additionally, much of the literature on indirect measures of assessment, such as knowledge surveys, is focused on their use by faculty as formative curriculum tools (Bell and Volckmann, 2001; Bowers et al., 2005; Nufher and Knipp, 2003; Wirth and Perkins, 2005). Thus, despite the benefits noted above, there is disagreement as to whether these indirect self-reported measures are valid indicators of student learning.

While Nuhfer and Knipp (2003) concluded that knowledge surveys, the indirect measure used in their research, provide a "detailed assessment of content, learning and levels of thinking" (p. 13), it is critical to keep in mind the context of their research. Nuhfer and Knipp (2003) were not comparing indirect measures to more direct assessments; their focus was on identifying a better measure of course success than using student evaluations. They concluded that knowledge surveys "improve planning and preparation, that they validate student learning much better than summative student ratings" (p.13). Most of their analysis evaluated knowledge surveys as a tool for curriculum design to identify course overlap and topics in need of additional instruction.

Other researchers demonstrated improvement in indirect measures but without extending their investigation to consider the relationship between indirect measures and other criteria of learning. Clauss and Geedy (2010) found increases in knowledge survey scores over the course of the semester. The emphasis of their research was on metacognition, and their analyses focused on the accuracy of the self-assessment measures rather than the relationship of self-assessment to other measures of student learning.

*Research on Direct and Indirect Measures*

Beyond an increase in scores over the course of a semester, what evidence exists to demonstrate that indirect measures accurately gauge student learning? A key methodological component to gathering such evidence is in the definition of student learning. As can be seen in the studies that follow, researchers define student learning in a variety of ways: 1) having students *answer* the indirect items, thus creating a direct measure of knowledge; 2) course grade; 3) exam grade; or 4) score on a direct measure unaffiliated with the course (locally developed or commercially available).

Wirth and Perkins (2005) administered a knowledge survey at several points during the semester noting increases in student confidence with course topics as well as a positive relationship between knowledge surveys and student performance. Knowledge surveys given before midterm exams correlated positively with midterm exam grade, $r = .70$, and a post course knowledge survey was similarly associated with final course grade, $r = .72$. Based on these findings, they suggested that indirect measures can accurately represent student knowledge. These correlations, however, may have been influenced by the fact that their knowledge survey was used in a formative manner; the survey was administered multiple times during the semester (prior to the course, prior to each exam, and after the course), used as a study guide by students, and had several items in common with the exams.

Bell and Volckmann (2011) demonstrated increases from pre to post knowledge surveys and found a positive correlation between knowledge surveys and final exam performance ($r = .58$). Their study thus, lends support to Wirth and Perkins (2005) that knowledge surveys can accurately assess student learning. However, Bell and Volckmann (2011) reported briefly on a third study that predated the focus of their 2011 article. In this pilot study, the knowledge survey was prepared by someone other than the instructor and results were not shared with the instructor or the students. They reported a substantially lower correlation between post course knowledge survey scores and final exams of $r = .37$ compared to $r = .58$ in the 2011 study when students had access to the knowledge survey throughout the course, and the final exam closely aligned with the knowledge survey items. This raises the question of whether the correlations observed between indirect and direct measures are in fact influenced by the use of the indirect measures as study guides throughout the semester and/or the fact that indirect items are often embedded in the direct measures.

Research by Price and Randall (2008) supports this conjecture, confirming increased levels in both direct and indirect measures of assessment across the semester. However, they failed to find a significant relationship between the two measures. Price and Randall (2008) demonstrated that students did not have the ability to differentiate between their perceived knowledge and their actual knowledge, and results supported the use of direct measures to assess student learning. The direct measure of knowledge used by Price and Randall (2008) was student responses to the indirect items taken at a later time in the semester; however, it is important to note that the indirect measures were used in more of a summative manner with neither students nor faculty having access to results or answers until the end of the semester.

Bowers et al. (2005) similarly found an increase in knowledge survey scores over the course of a semester, but concluded that indirect measures are not accurate estimates of student knowledge. The magnitude of the correlation of indirect score with final grade was quite variable over the five sections of the Biology course under study (ranging from $r = .21$ to $r = .46$). Bowers et al. (2005) additionally analyzed item pairs between the indirect measure and the final exam, and

found little agreement. Thus, the literature provides discrepant findings when comparing indirect and direct measures.

Although prior studies of indirect knowledge, (Bell and Volckmann, 2011; Bowers et al., 2005; Clauss and Geedney, 2010; Nufher and Knipp, 2003, Price and Randall, 2008; Wirth and Perkins, 2005) all provided evidence of increases from pretest to posttest demonstrating increased familiarity with course topics, when indirect measures were compared to direct measures, results were mixed. Findings are further confounded by variations in the use of indirect measures as formative or summative tools. In several studies indirect measures were used as study guides throughout the semester, and often the same indirect items appeared on later direct measures of student knowledge creating a situation that would inflate the relationship of these measures. Additionally, the motivational level of respondents differed across studies with some test scores contributing to course standing/grades (Price and Randall, 2008), others providing minimal credit for participation (Bowers et al., 2005), and still others varying in the confidentiality of survey results among students and faculty. To further complicate the findings, several researchers suggested that the accuracy of self-assessment is not consistent across students and this inconsistency follows a specified pattern (Bell and Volckmann, 2011).

### *Dunning-Kruger Effect*

Kruger and Dunning (1999) were among the first researchers to systematically study a metacognitive deficiency where individuals are unable to accurately self-assess their performance on various tasks. Through a series of four studies spanning humor, grammar, and logical reasoning tasks, they demonstrated an overestimation of self-ability by the lowest performers and underestimation by the highest performers. These miscalibrations were disproportionately large for the lowest performers, suggesting more than a regression effect as an explanation. Through extensive experimentation, Kruger and Dunning (1999) were able to demonstrate that lower performers could not learn to recalibrate their own ability by observing the superior performance of others, but could achieve increases in accuracy if trained in the specific task being assessed (in this case logical reasoning). Kruger and Dunning's (1999) findings appear at odds with Bell and Volckmann (2011), as one would expect that by the end of a course sufficient training and feedback on content would lead to greater accuracy in self-assessment.

An application of these findings to academic assessment can be found in the work of Bell and Volckmann (2011). These researchers found correlations between knowledge survey scores and final exam performance ($r$ = .58), but the relationship was inconsistent across student ability levels. Students who scored lower on the final exam were overconfident in their estimated ability (most dramatically in the post-measure of indirect knowledge, but also relatively overconfident in a pre-measure as well). Thus, the very students faculty would hope to assist with the use of an indirect measure – those low in ability – may not benefit due to the inaccuracy of their self-assessment.

Hacker, Bol, Horgan, and Rakow (2000) confirmed the inaccuracies in the lowest performers in their study of undergraduates in an educational psychology course. Their course incorporated instruction on the benefits of self-assessment and provided opportunities for students to self-assess before and after each exam. They showed that overall accuracy increased over time with repeated self-assessments; but not for the lowest performers.

## Current Study

Previous literature has focused on examining indirect measures and direct measures of student learning in science and quantitative courses. Price and Randall (2008) investigated a quantitative course in the business major; Bell and Volckmann (2011) a general chemistry course; Wirth and Perkins (2005) a geology course; and both Clauss and Greedey (2010) and Bowers et al. (2005) researched biology courses. This study is one of the first to address a course in the social sciences, a core psychology research methodology course. Hypotheses 1, 2, and 3 propose to confirm results from earlier research with a focus on a social science research course.

$H_1$: Average posttest indirect scores will be higher than average pretest indirect scores.

$H_2$: Average posttest direct scores will be higher than average pretest direct scores.

$H_3$: Average posttest direct scores will correlate positively with final course grades.

Several earlier studies used the indirect measures in a more formative manner allowing students and faculty access to the measure throughout the semester. This resulted in benefits of providing feedback and allowing for changes in study or course focus, but, such a practice may have exaggerated the relationship between indirect and direct measures. Our focus is on the use of indirect measures as a purely summative tool. Thus, while employing a pretest/posttest methodology for administration, the students' performance was not shared with students or faculty. Nor were faculty aware of the assessment items except to the extent that the items mirrored formal course descriptions and learning objectives.

Other studies employed faculty teaching the course to assist in the development of the measures, however, in our study both measures were developed by a senior faculty member who was not teaching the course. Thus, a broader view of the learning outcomes was employed, avoiding bias toward an individual faculty member's pedagogical perspective or course focus.

We also differed from earlier research in that our indirect measure was Likert scale responses to course terms or concepts, not to actual direct knowledge items. For example, other researchers assessed indirectly by repeating the direct knowledge item and having respondents self-report their level of confidence or ability to solve the problem/answer the question. We, however, had respondents read course terms or concepts and indicate their familiarity/confidence level with each item. Thus, the indirect responses were more general and not a response to a specific test item. We paired indirect items and direct items to ensure each item was represented on both measures *(Figure 1).*

| | Indirect Items | Direct Items |
|---|---|---|
| Development | Development was based on course topics identified through a review of course outlines and syllabi, followed with a review by faculty. | Knowledge or application questions were created that corresponded to the course topics. |
| Response Format | Students answered indirect items on a 4-point Likert scale. | Students chose the correct response for direct items using a multiple choice format. |
| Example *(Items were not presented as* | "Reliability in Measurement"<br>1. Have never heard of this. | Dr. Archibald and Dr. Santos appeared as expert witnesses for opposing sides in a murder trial. They |

| | | |
|---|---|---|
| *pairs in the instrument)* | 2. Have heard of this, but don't really know what it means.<br>3. Have some idea what this means, but not too clear.<br>4. Have a clear idea of what this means and can explain it. | had each independently conducted a psychological evaluation of the accused to determine his competency to stand trial. The independent reports by the two doctors stated that the accused was competent. The evaluation by these two experts demonstrated:<br><br>a. Generalizability<br>b. Reliability<br>c. Standardization<br>d. Validity |

*Figure 1.* **Indirect and Direct Item Example** Paired indirect item and direct item for the course topic of "reliability in measurement".

Due to the lack of availability of the test items to students and faculty, the use of "paired" items rather than identical items across the indirect and direct measures, the more general focus of both measures, and the potential differences in student motivation, we offer hypotheses in support of Price and Randall (2008), and Bowers et al. (2005).

$H_4$: Posttest indirect scores will fail to correlate positively with posttest direct scores.

$H_5$: Posttest indirect scores will fail to correlate positively with final course grades.

Prior studies disagree as to both the existence and the cause of the Dunning-Kruger Effect.

Given the differences already cited between this study and other researchers (the use of a measure as a summative tool, the general nature of indirect items, test items, and the results not accessible by faculty) we do not propose a specific result but will simply test if students vary in their ability to accurately self-assess.

$H_6$: Respondents low in ability (measured as direct post score) will differ in their self-assessment of knowledge and abilities (measured as indirect post score) compared to respondents high in ability.

## Method

### *Participants*

Students enrolled in six sections of an undergraduate Psychology research methods course at a mid-size 4-year public college in the northeast United States during the fall 2011 semester were invited to participate in both the pretest assessment and the posttest assessment. The pretest, which includes both the indirect assessment and the direct assessment was administered on the first day of classes to all those enrolled, and most students participated for an initial response rate of 88%. The same instrument was administered as a posttest at a date chosen by the instructor but within the last two weeks of the semester. Of the 108 participants who took both the pretest and posttest, a total of 38% failed a validity check item (34 participants pretest, 13 participants posttest), and were eliminated. The validity check was an instructed response item embedded in the indirect assessment and was of the format "For validation purposes, please respond 'Strongly Disagree' for this item". Analysis was performed on the remaining 67 participants.

Gender and ethnicity were not collected, however data on year in college and status as a Psychology major were. Respondents were 22% freshmen, 37% sophomores, 28% juniors, and 12% seniors. The majority of participants were psychology majors or double majors, 30%, and an additional 34% stating that they planned to apply to the major. In terms of being a psychology minor, 18% were declared as such with an additional 12% indicating that they planned to apply to the minor. Finally 6% were not affiliated with psychology as a course of study.

*Materials and Procedures*

The psychology research methods course is the second prerequisite course in a series of four courses in the psychology major's methodological core, and is a requirement for higher level courses in the major. This course is also a requirement for the psychology minor (The College of New Jersey Psychology Department, 2015). Formal course descriptions and syllabi from the previous academic year were used to identify 55 subtopics that were relevant for the course. An on-line survey was developed where the current and recent faculty teaching the course rated each topic as "essential," "optional," or "unessential" for inclusion in the course. A senior faculty member who was not currently teaching the course developed test items to measure the subtopics rated as "essential" by at least 50% of the faculty respondents. The faculty teaching the course were only involved in reviewing the general topics, and did not see the final test items.

Indirect items were created in the manner of Cross and Angelo (1988) as modified by Eder (2009). These items were developed to measure familiarity with terms/concepts of the research methods course (using the same list of subtopics identified above by current faculty). Unlike other indirect measures of knowledge these items presented course terms and concepts rather than actual exam questions. A total of 49 concepts such as "convenience sampling," "independent variable," and "ordinal level of measurement," were presented along with a 4-point Likert scale to indicate familiarity with the topic *(Figure 1)*. Relative to other studies, ours differed in that the direct measure was created, administered, and evaluated by an objective 3rd party faculty member. Additionally, faculty teaching the course did not have access to or knowledge of the instrument itself or their course section's pretest or posttest scores until a report was compiled for the department a full year later. Even at that time, data were only available in the aggregate, not by section or by instructor. Thus, our use of the indirect measure was more summative than formative. We sought to determine if the indirect measure, which is simpler to construct and administer, could serve as a "stand alone" measure of student learning. If a knowledge survey or indirect measure could accurately represent actual knowledge gained, we could use these tools in future assessments. The 42 direct knowledge items were all objective, presented in a multiple choice format with a single correct answer. Thus, higher scores indicated greater knowledge of the course content.

The two measures were combined into a single on-line survey which took about 25 minutes to complete. The survey was administered in a pretest/posttest fashion in the fall 2011 semester. Only those respondents who completed both the pretest and the posttest, and passed a validity check were retained, resulting in 67 participants. Cronbach's coefficient alpha was used to determine reliability of the measures resulting in $\alpha = .75$ and $\alpha = .92$, for direct and indirect, respectively.

Final course grades were obtained from the student record system and could range from F through A. These were converted to a numerical value (A = 11, A- = 10, B+ = 9, B = 8, B- = 7,

C+ = 6, C = 5, C- = 4, D+ = 3, D = 2, F = 1). Because the resulting grade was ordinal, Spearman correlations were used to test Hypotheses 3 and 5.

## Results

Prior to conducting the analyses, a series of statistical checks were run to ensure that the responses we eliminated due to a failed validity check did not distort the findings. Failing a validity check is not an unexpected occurrence in low stakes assessments, especially on a pretest measure, when respondents have little knowledge of the test content. However, to ensure that the sample of respondents who were removed did not differ from the final sample used in the analyses, several statistical tests were conducted. A series of *t*-tests found no differences between these two groups for any of the measures – pretest indirect, posttest indirect, pretest direct, or posttest direct. The measure of final grade more reflected the properties of ordinal data than interval level, thus this variable was analyzed using chi-square and again, no differences were found in the distribution of grades for these two samples. This was true whether analyzing all levels of final grade or final grade recoded into low, medium, and high groups.

The grade distribution was negatively skewed as 46% of students received a grade of A. Again, a check of the eliminated responses against the 67 responses retained for analysis was conducted to ensure the removal of these cases did not skew the data. Analyses of all measures revealed that the removal of those who failed the validity check did not impact the distribution of any of the measures. With or without those cases, the data reveal a strong negative skew evident in course grades and in posttest indirect scores for both groups. This high percent was expected and mirrors the recent trend of grade inflation in higher education, reported as 43% of letter grades being A (Rampell, 2011).

Hypothesis 1 examined the difference between pretest and posttest average indirect scores by using a paired samples *t* test to calculate the mean difference between scores across time. Referring to Figure 2, posttest indirect scores ($M$ = 163.30, $SD$ = 15.22) were statistically significantly higher than pretest indirect scores ($M$ = 123.73, $SD$ = 17.37), $t(66) = 14.56$, $p < .001$, $d = 2.42$, $CI_{.95} = 44.99, 34.14$ supporting Hypothesis 1.

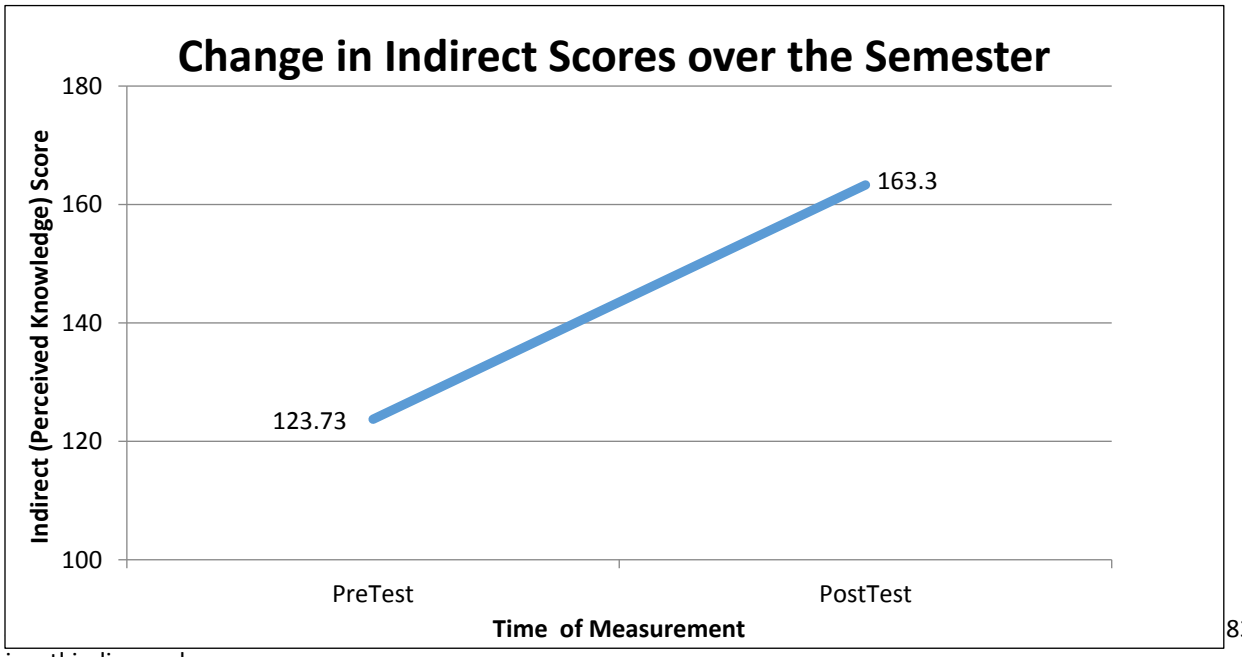

*Figure 2*. **Change in Indirect Scores over the Semester** Mean change pretest to posttest indirect measures $t(66) = 14.56$, $p < .001$, $d = 2.42$, $CI_{.95} = 44.99, 34.14$ resulting in higher posttest indirect scores supporting Hypothesis 1.

Hypothesis 2 examined the difference between pretest and posttest average direct scores by using a paired samples *t* test to calculate the mean difference between direct measure scores across time. Referring to Figure 3, posttest direct scores ($M = 36.69$, $SD = 5.44$) were statistically significantly higher than pretest direct scores ($M = 29.50$, $SD = 5.60$), $t(66) = 11.31$, $p < .001$, $d = 1.30$, $CI_{.95} = 8.46, 5.92$ supporting Hypothesis 2.

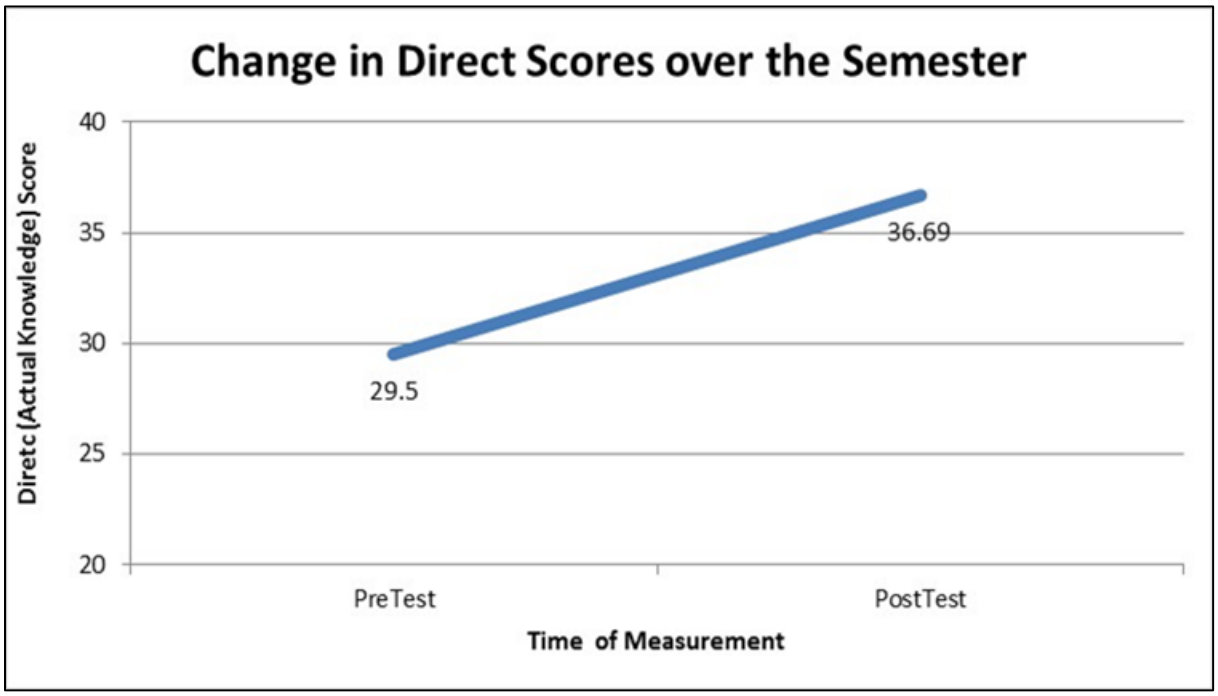

*Figure 3.* **Change in Direct Scores over the Semester** Mean change pretest to posttest direct measures $t(66) = 11.31$, $p < .001$, $d = 1.30$, $CI_{.95} = 8.46, 5.92$ resulting in higher posttest direct measure scores supporting Hypothesis 2.

Hypothesis 3 examined the relationship between posttest direct scores and final course grades, and was addressed by using Spearman correlation. A statistically significant relationship was found between posttest direct scores and final grades, $r_s(67) = .47$, $p < .001$. Grades correlated in a positive direction with posttest direct scores, therefore Hypothesis 3 was supported.

Hypothesis 4 examined the relationship between posttest indirect scores and posttest direct scores, and was addressed by using Pearson correlation. No statistically significant relationship was found between posttest indirect scores and posttest direct scores, $r(67) = .11$, $p = .357$. Students did not demonstrate the ability to self-assess their knowledge, and scores did not correlate significantly in a positive direction, therefore Hypothesis 4 was supported.

Hypothesis 5 examined the relationship between posttest indirect scores and final course grades, and was addressed by using Spearman correlation. No statistically significant relationship was found between posttest indirect scores and final grades, $r_s(67) = .06$, $p = .615$. Students did not demonstrate the ability to self-assess their perceived knowledge, and grades had no relationship with posttest indirect scores, therefore Hypothesis 5 was supported.

Hypothesis 6 examined if direct measure low scorers differed in self-assessment of their knowledge and abilities than high scorers. This was addressed using a mixed-model ANOVA where the between-subjects variable was the three levels of direct knowledge and the within-subject variable was the measure of indirect knowledge measured at the beginning and ending of the course. Referring to Figure 4, direct scores were divided into three intervals comprising the bottom, the middle, and the top third of the frequency distribution. A statistically significant difference was found between pretest ($M = 123.73$, $SD = 17.38$) and posttest ($M = 163.30$, $SD = 15.22$) indirect scores, $F(1, 64) = 205.30$, $p < .001$, $\eta^2 = .762$. However, there was no significant interaction of pretest and posttest scores, $F(2, 64) = 0.11$, $p = .896$, $\eta^2 = .003$, across the three ability levels, suggesting that the within group difference over time was similar for all three groups. There was a statistically significant difference for indirect scores across the three levels of direct knowledge: low ($M = 142.93$, $SD = 18.08$), moderate ($M = 138.65$, $SD = 15.77$) and high ($M = 148.91$, $SD = 13.70$), $F(2, 64) = 4.76$, $p = .012$, $\eta^2 = .13$. A post-hoc comparison of means using Tukey's *HSD* at $\alpha = .05$ revealed that the moderate direct knowledge group was significantly lower in perceived ability than the high direct knowledge group. The low direct knowledge group did not differ from the moderate or high direct knowledge groups as shown in Figure 4.

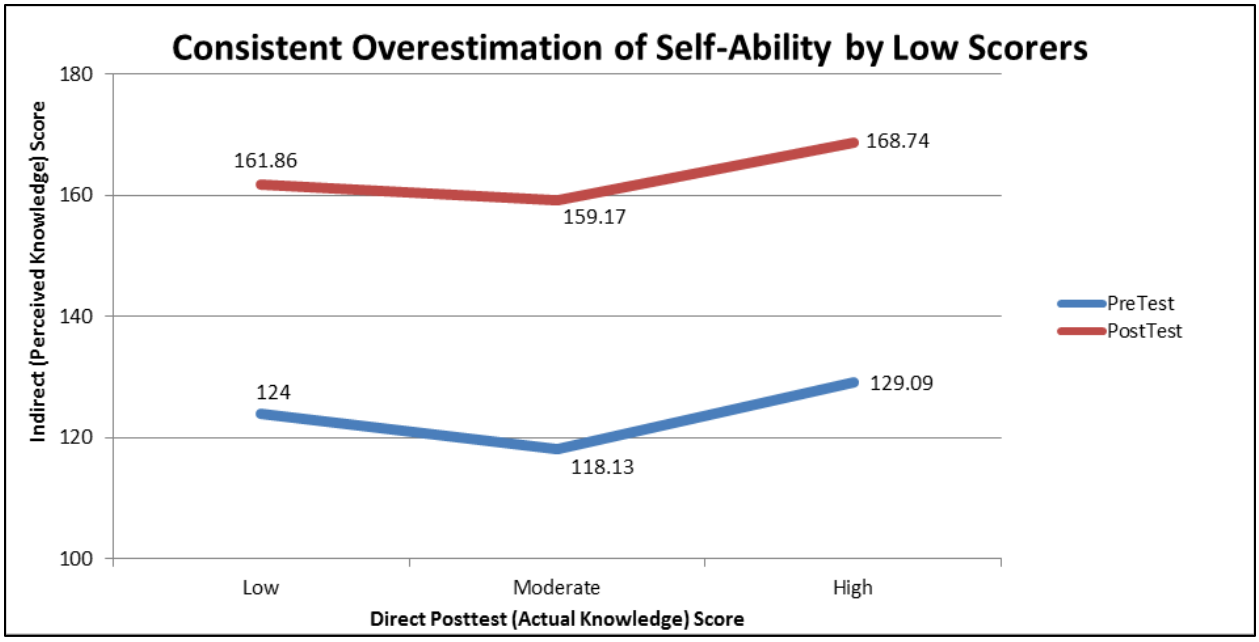


*Figure 4.* **Consistent Overestimation of Self-Ability by Low Scorers** The moderate direct knowledge group was significantly lower in perceived ability than the high direct knowledge group. The low direct knowledge group did not differ from the moderate or high direct knowledge groups.

## Discussion

Two different measures of assessing student learning were administered at two points in time to six sections of a psychology research methods course at a higher education institution. Indirect and direct assessments were used to determine levels of student learning by measuring both perceived knowledge and actual knowledge of course content. The indirect assessment measured students' confidence levels on their knowledge of course topics, and was further used to determine if an indirect measure could accurately gauge student learning. The direct assessment addressed student actual knowledge with items that paralleled the indirect course topic items; results were examined along with final course grades to determine if students acquired the essential knowledge and skills from the goals and objectives contained in the course curriculum.

We found the level of perceived knowledge from our indirect measure increased significantly from pretest to posttest across all sections of the course, indicating increased confidence and familiarity with course topics over time. This finding supports prior research (Bell and Volckmann, 2011; Bowers et al., 2005; Clauss and Geedey, 2010; Nuhfer and Knipp, 2003; Price and Randall, 2008) where student confidence levels showed an increase by the end of the semester. Similarly, the level of actual knowledge from the direct measure increased significantly from pretest to posttest, demonstrating students completed the course with more knowledge than they had at the beginning of the semester. In addition to statistical significance, both measures demonstrated a strong effect size, further supporting the magnitude of these differences, or gains in knowledge. Our results support Bell and Volckmann (2011) and Price and Randall (2008) who used pretest and posttest direct measures and showed gains in actual knowledge of course content by the end of the semester. Increases in pretest to posttest scores can identify areas in need of improvement and determine if the department and the faculty are "adding any value" demonstrated

by increased knowledge in students (Pederson and White, 2011). Although this study, along with prior research found increased scores from pretest to posttest using indirect perceived knowledge measures and direct knowledge measures, the results do not support the use of indirect measures as indicative of student learning.

This study found indirect measures of perceived knowledge were not an accurate indicator of student learning, as our indirect measure did not correlate with the direct knowledge measure or course grades. Our results support prior studies (Bowers et al., 2005; Price and Randall, 2008) in which students were unable to accurately perceive their knowledge and abilities via indirect measures. Bell and Volckmann (2011) and Wirth and Perkins (2005) on the other hand, found indirect posttest measures of perceived knowledge a good predictor of later performance on a test. However, when the indirect assessment was compared to final course grades, mixed results were found. In accordance with Bowers et al. (2005) results from the indirect perceived knowledge measure were not a good indicator of final course grades. Wirth and Perkins (2005) however, differed from these results, and found perceived knowledge a good indicator of final course grades.

Unlike previous research, we did not provide the indirect assessment to students as a study guide; in the other studies, this may have aided students in posttest preparation. Our small correlation may also be due to a lack of course consistency across the six sections of the psychology research course. Five of the six instructors who taught the course in the fall 2011 semester were adjunct faculty, and three of them were teaching the course for the first time. In developing our assessment, we consulted with faculty teaching the course, and created indirect items and direct items based on topics deemed essential, however individual instructor variation across six sections of the same course can be difficult to control without a more uniform course design. A possible solution to ensure course consistency in future assessments is to involve all faculty members registered to teach the course in the design and the implementation of the assessment measures.

The skewed distribution of final course grades may have weakened the correlation between the indirect scores and final course grades, due to restriction in range. Future researchers might minimize the impact on correlations by obtaining a measure of final course grade that is numerical (on a basis of 0 to 100) as opposed to a graded option (F through A). While skewness would still be evident, the finer measurement might mitigate some of the effect of range restriction. It is important to note however, that any restriction in final course grade would have similarly affected the direct measure's relationship with course grade, which was statistically significant. It seems reasonable to assume that if the correlations were stronger in a less restrictive sample, the magnitude of the difference in the correlations for direct and indirect score would still exist. Thus, the conclusion that the indirect score is not as effective as a measure of course knowledge is still supported.

Lastly, based on prior research (Bell and Volckmann, 2011), this study attempted to determine if low scorers on the posttest direct measure were more confident in their knowledge and ability than high scorers. In accordance, we found low direct measure scorers overconfident in their abilities (measured by the direct score). Bell and Volckmann (2011) suggested overconfident low scorers were a result of their inability to self-assess accurately. Clauss and Geedey (2010) suggested students with excellent self-assessment skills may do poorly on the measure, then use it as a study guide and do better on the direct measure thus, explaining some of the disconnect between the two measures as maturation/developmental. However, this rationale would only apply in studies that use the indirect measure in a more formative manner.

Another explanation may exist if the direct measure is not an accurate portrayal of student knowledge. Could the scores on the direct measure be negatively influenced for the low scoring students by test anxiety or difficulty with the type of measurement – a multiple choice objective test? As a test of this, one might measure test anxiety and control for this in future studies. Final grade information could be obtained separately for homework, projects, tests, and quizzes to see if the correlations vary depending on the format in which student knowledge is measured.

One might ask the question: "Could the inability to accurately self-assess be a stable and constant occurrence?" The respondents in this study were over confident in the pretest measure as well as the posttest measure. One might postulate that the posttest measure should have had a self-correction factor as students were receiving grades and participating in course assessments with feedback throughout the semester. Kruger and Dunning (1999) suggested that increasing one's competence in the knowledge/ability being assessed was a method to improve self-assessment. However, despite a semester long process of learning, testing, and feedback, we still found discrepancies in self-assessment accuracy. It would be interesting to measure student self-assessment pretest and posttest across several different courses or tasks and determine if this is a stable trait or consistent self-report bias.

Mabe and West (1982) identified both person (i.e., internal locus of control, high achievement status, high intelligence) and measurement conditions (anonymity, expected comparison of self-assessment with a criterion measure, prior experience with self-assessment, self-evaluation directive that emphasizes social comparison to others) that were predictive of ability to self-assess. While many of these variables were present in the current study and prior research, they most often serve as confounding elements and not as controlled independent variables. Future studies might investigate these and other variables in a systematic and controlled methodology.

Approaching the use of these measures with a summative focus also may have influenced the motivation of our participants. On the negative side, motivation is removed as our measures did not "count" towards a grade, nor did we provide direct feedback to the student. However, the confidentiality of scores also removes a need for social desirability and impression management (which could have artificially raised the indirect scores) that students might engage in as a response to the pressure that their instructor will see the results.

Several aspects of our sample contribute to the limitations of this study. Although statistical analyses ensured that responses removed for failing a validity check did not differ on key measures from those who passed the validity check, the high percent of respondents failing is still concerning. We have moved to a system where validity checks are now embedded throughout department assessments so that they might serve not only as detection, but as early warning and prevention. As with any small scale research project, these results may not generalize to other locally developed measures or other settings. Our small sample size reduced power in our analyses, compromising our ability to demonstrate statistical significance. However, the importance of research on local measures cannot be overstated. Indeed, a recent survey of 1,200 degree granting institutions of higher learning in the United States revealed that almost 50% reported using locally developed measures of student knowledge and skills (Kuh, Jankowski, Ikenberry, and Kinzie, 2014). While locally developed measures and findings may not generalize across institutions, the need for continued conversation is evident.

This study demonstrated the utility of a direct measure of student learning, but not that of an indirect measure of student learning. Although this study's indirect measure was not an accurate indicator of student learning, and students who were low performers systematically overestimated

their abilities, the use of such measures may still have future benefits. Bowers et al. (2008) found evidence from faculty that students used indirect measures to prepare for exams. Nufher and Knipp (2003) and Price and Randall (2008) found that students provided with the indirect knowledge measure can identify areas of weakness in course topics when results are analyzed at the item level. Thus, while not a substitute for direct measures as assessment tools, indirect measures may play a role in clarifying expectations of course content and student preparedness while also serving as a guide for progress during the semester. Further, when learning outcomes are not concerned with course knowledge, but rather attitudes, perspectives, and expectations, indirect measures may be optimal. However, in situations where the goal of assessment is to demonstrate course knowledge in a summative manner, direct measures are clearly superior.

## Acknowledgements

The authors wish to acknowledge the editorial staff at the *Journal of the Scholarship of Teaching and Learning* and technical reviewers at Educational Testing Service for their suggestions on an earlier version of this article, the Psychology faculty at The College of New Jersey (TCNJ) for allowing access to classes, and the TAPLab (Testing and Assessment in Psychology) at TCNJ for assistance in data collection.

## References

Banta, T. W. (2004). Developing assessment methods at classroom, unit, and university-wide levels. Paper prepared for the Scottish Higher Education Agency. Retrieved October 30, 2012 from: http://www.bmcc.cuny.edu/iresearch/upload/Banta.pdf

Bell, P., and Volckmann, D. (2011). Knowledge surveys in general chemistry: Confidence, overconfidence, and performance. *Journal of Chemical Education*, *88*(11), 1469-1476.

Bowers, N., Brandon, M., and Hill, C. (2005). The use of a knowledge survey as an indicator of student learning in an introductory biology course. *CBE Life Sciences Education*, *4*(4), 311-322. doi:10.1187/cbe.04-11-0056.

Clauss, J., and Geedey, K. (2010). Knowledge surveys: Students' ability to self-assess. *The Journal of Scholarship of Teaching and Learning*, *10*(2), 14-24.

Cross, K. P., and Angelo, T. A., (1988). Classroom assessment techniques: A handbook for faculty. Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, University of Michigan.

Eder, D. (2010, October). "Closing the loop: No money, no time, and I'm expected to assess, too?" Presented at the Assessment Institute, Indianapolis IN.

Fort, A. O. (2011). Learning about learning. *Liberal Education, (Winter)* 56-60.

Hacker, D.J., Bol, L., Horgan, D. D., and Rakow, A. R. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160-170.

Huffman, L., Adamapoulos, A., Merdock, G., Cole, A. and McDermid, R. (2011). Strategies to motivate students for program assessment. *Educational Assessment, 16*, 90-103.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., and Kinzie, J. (2014). Knowing what students know and can do: The current state of student learning outcomes assessment in US colleges and universities. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Liu, O. L., Bridgeman, B. and Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*(9), 352-362.

Mabe, P. A. and West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280-296. doi: 10.1037/0021-9010.67.3.280

Martell, K. (2007). Assessing student learning: Are business schools making the grade? *Journal of Education for Business, 82*(4),189-195.

Nuhfer, E., and Knipp, D. (2003) The knowledge survey: A tool for all reasons. Retrieved October 31, 2015 from: http://pachyderm.cdl.edu/elixr-stories/resource-documents/knowledge-survey/KS_a_too_for_all_reasons.pdf

Office of the Provost (2015). *Outcomes Assessment.* University of Wisconsin-Madison. Retrieved October 31, 2015 from: http://www.provost.wisc.edu/assessment/manual/manual2.html#a3

Pedersen, D. E., and White, F. (2011). Using a value-added approach to assess the sociology major. *Teaching Sociology*, *39*(2), 138-149.

Price, B. A., and Randall, C. H. (2008). Assessing learning outcomes in quantitative courses: Using embedded questions for direct assessment. *Journal of Education for Business*, *83*(5), 288-294.

Rampel, C. (2011, July 14). A history of college grade inflation. *The New York Times.* Retrieved October 31, 2015 from: http://economix.blogs.nytimes.com/2011/07/14/the-history-of-college-grade-inflation/

The College of New Jersey. (2015). Psychology Major & Specializations. Retrieved October 31, 2015 from: http://psychology.pages.tcnj.edu/academic-programs/psychology-major-specializations/

Wirth, K.R., and Perkins, D. (2005). Knowledge surveys: An indispensable course design and assessment tool. *Innovations in the Scholarship of Teaching and Learning*. Retrieved from www.macalester.edu/geology/wirth/WirthPerkinsKS.pdf