Measurement Usage in Self-Assessment

Yaelan Wong[1], John Lovell[2], John Norton[3], and Julia A. Norton[4]
[1]MS student, Department of Statistics and Biostatistics, [2]Department of Psychology, [3]BS student, Department of Mathematics and Computer Science, [4]Department of Statistics and Biostatistics, California State University, East Bay, Hayward, CA 94542

Abstract
Continuing our study of self-assessment of individual exam questions, we turn to various forms of rules for assigning values to the question 'How certain are you of the correctness of your answer?' In several introductory psychology classes we measured self-assessment as it correlates to correct response or incorrect response using several rules for assigning assessments. When we originally used a 1 to 5 scale (1 certain correct and 5 no knowledge), we noticed that some individuals always responded sure or unsure (1 or 5) with no further gradation. These results generally corresponded to better students using a 1 rating and poorer students using a 5 rating. Two additional rules were tested. Where possible we have repeated tests over individuals using the same rule. Our results are compared with past results with these new rules or instructions given to students during testing.

Key Words: Educational assessment; Learning; Teaching

1. Introduction

Trends in assessment have turned lately to asking students how well they learned the material in a course. As with most publicly funded state colleges and universities, there is a continued effort to do more with less funding, to combine, reduce or eliminate small programs, and to demand that all programs demonstrate that they provide something for the public good. Dietz, Lovell and Norton have collaborated on a number of issues in educational assessment over the past ten years including learning in introductory psychology and statistics courses. (2000, 2005, 2012)

Summaries of our previous studies appear in several *Proceedings of the American Statistical Association Section on Statistical Education* as our data increased and the questions became more varied (Lovell, Dietz, Eudey and Norton with others between 2000 and 2006). These papers consider assessments in introductory courses and our statistics degree programs. The ideas discussed are consistent with the fundamental learning goals outlined in Garfield and Ben-Zvi (2007) and Norton and Lovell (1981). In 2006-2007 Norton served as Interim Director of Institutional Research, writing a broader survey of the assessment at the University (Norton 2007). Returning to teaching in 2007-2008 Norton collaborated with many faculty from all areas of the university in supporting assessment attempts (Norton, Zhou, and Ganjeizadeh 2008 and Eudey, Anand, Norton and Coulman 2009).

Seeking less controversial means of evaluation among university faculty and ones perhaps less intrusive into the classroom, some suggest asking students directly about their learning experience in terms of what they had learned using a consumer model of assessment. Since our previous research concluded that common finals written by committee or by outside evaluators gave results that satisfied us, we wondered how a

version of these new methods might work. We are not in favour of using student evaluations as assessments. Therefore, in 2011 we decided to associate the question of learning with the twenty questions already being used in the introductory statistics final. Achieving some indication of student ability to determine correctly how difficult a particular question was for an individual to answer, we decided to extend our investigation to introductory psychology courses as well. The results were mixed. In 2012 and 2013 we included similar results from an Introductory Psychology course to see how the results varied.

2. Relationship Between Correct Response And Student Certainty

We wondered whether the correct and incorrect responses related to the degree to which students were certain of their answers (Dietz, Lovell, Norton and Norton 2012). For each question on the course final, students indicated on a scale from 1 to 5 how certain they were of the answer that they had given in that work. The scale was ranked from highest to lowest. Indicating 1 meant that the student was very sure that the response given was correct, while indicating 5 meant that the student was very unsure of the given response. The value 3 represented neutral on this scale, neither sure nor unsure about the response. Certainty of response and correct response were associated. . We expected better students to have correlations, but we were unsure whether this was indeed realistic. These results were verified in 2012. Would a different measurement scale give stronger results?

In several introductory Psychology courses, Lovell assigned students to order the problems in difficulty from 1through 20. That experiment was not worth the students' time and few actually completed the task. In several other courses students were assigned to indicate on a scale from 1 to 10 how certain they were of the answer given for each problem. We have data for the same students over three exams and also data from the final for more terms. The two main questions for these results are (1) the point-biserial correlations between whether a problem is correct or not and the level of assuredness a student says she/he has for that answer on the final is stronger for the five or ten point scale and (2) do students improve in evaluating their abilities over a term?

Our subjects are 307 students in four introductory psychology classes at California State University East Bay, all taught by the same instructor and given credit for carefully completing the self-assessment as they completed one exam. We use the measure point-biserial correlation coefficient between correctly responding to a question and the perceived difficulty as reported by each student and its transformation using Fisher's z statistic for correlation. To ensure some degree of similarity, the exam scores are compared across the three assessment methods as well. The exam scores do not significantly differ by method as they should not. There is some problem of confounding as methods are nested in classes and really these two effects can't be distinguished. The result that exam scores do not differ by method is reassuring at least. Looking at Studentized residuals and deleted residuals for the variables shows a poor fit for test scores, but are markedly uninteresting for the three other variables.

Table 1 below shows the results for repeated simple analysis of variance to the compare the test scores (%Acquisition) over the three measurement schemes as well as the point-biserial correlations and their Fisher Z-transformations, scaled and unscaled for differing numbers of questions, again against the three measurement schemes. Among the three measures, comparisons indicate that only for test scores (%Acquisition) are there no differences ($F=2,56$, $p<.079$). No matter how one compares the three measurement

schemes (using point-biserial correlations, or Fisher transforms), the methods of ranking, five-point scale and 10-point scales are not the same. Figure 1 shows means and pattern of differences in means. The usual tests results for all three measures are reflected in Figure 1.

ANOVA comparing measures, ranking, 5 scale, 10 scale

	Sum of Squares		df	Mean Square	F	Sig.
% Acquisition	Between Groups	1230.6	2	615.3	2.56	.079
	Within Groups	73854.2	307	240.6		
	Total	75084.8	309			
Point-biserial r	Between Groups	0.9	2	.43	8.94	.000
	Within Groups	14.3	294	.05		
	Total	15.2	296			
Fisher's Z unscaled	Between Groups	2.2	2	1.1	8.69	.000
	Within Groups	36.9	294	.13		
	Total	39.1	296			
Fisher's z scaled for n	Between Groups	15.6	2	7.8	5.36	.005
	Within Groups	426.9	294	1.4		
	Total	442.5	296			

Table 1. Repeated one-way ANOVA comparing the 3 scales on four variables (multivariate results similar and shown later). There are no differences on test scores and all three self-assessment measures show similar results, namely that the ten point scale is the best of the three scales used. Multivariate analysis and robust tests gave similar results. See Figure 1 for pattern of means. Only results for point-biserial are shown.

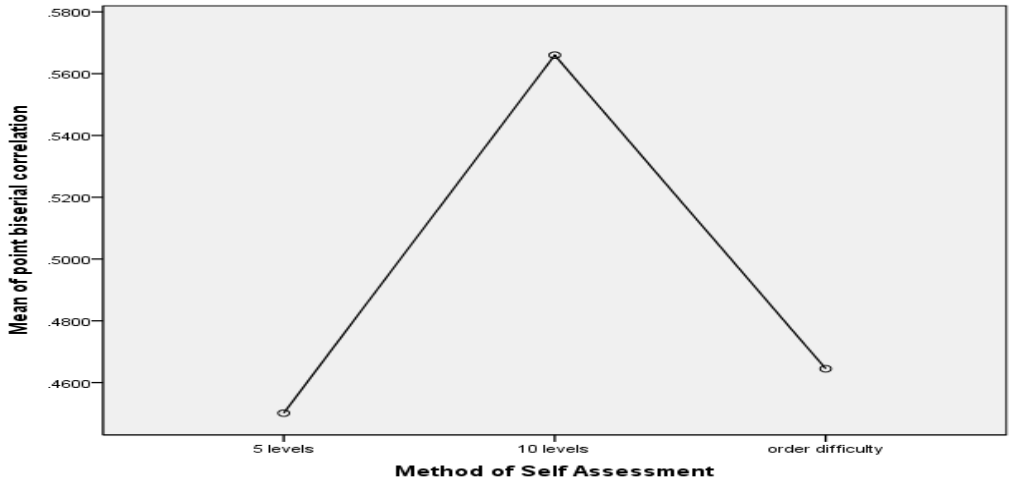

Figure 1: The measure point biserial correlation compares the three methods of self-evaluation, evaluation of difficulty using: 5 levels of a Likert scale, 10 levels of a Likert scale, or by ranking the difficulty of each problem ($F_{2,294} = 8.94$, $p < 0.001$).

3. Does learning occur as self-assessment is repeated over a term?

We study the 10 point Likert scale where repeated self-assessments are made within subjects over 2 and 3 exams. There is little evidence that students learn this self-assessment technique. Fisher's Z statistic transformation for the correlation gives no support for learning when looking at all 3 tests where they exist or for the first two. When the raw point-biserial coefficient is used as the summary measure of self-assessment, the subjects who have three exams show a quadratic relationship from the first test to the final exam. It is unclear why this should be. Comparing the test scores from the first to the final exam shows definite learning patterns from the first exam to the second to the final exam. The pattern is not the same as that for the correlation between what a student perceived as learned and what was actually been retained from the course. In any event the results are not strong. Only the point-biserial and the test acquisition results are included here. The other measures show similar patterns.

Point-biserial measures indicating learning of self-assessment technique
Measure: Point-biserial Correlation **Within-Subjects Factors**

Learning	Dependent Variable
1	posr2 point-biserial correlation exam 1
2	posr3- point-biserial correlation exam 2
3	Posrf- point-biserial correlation final exam

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Learning	Pillai's Trace	.174	4.014[a]	2.0	38.0	.026
	Wilks' Lambda	.826	4.014[a]	2.0	38.0	.026
	Hotelling Trace	.211	4.014[a]	2.0	38.0	.026
	Roy's Largest	.211	4.014[a]	2.0	38.0	.026

a. Exact statistic b. Design: Intercept Within Subjects Design: Learning

Table 4. Repeated measures model for the 3 exams indicating learning of self-assessment techniques. See Figure 5 for pattern of learning.

Tests of Within-Subjects Effects Measure: Point-biserial Correlation

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Learning	Sphericity Assumed	.493	2	.247	4.583	.013
	Greenhouse-Geisser	.493	1.957	.252	4.583	.014
	Huynh-Feldt	.493	2.000	.247	4.583	.013
	Lower-bound	.493	1.000	.493	4.583	.039
Error (Learning)	Sphericity Assumed	4.199	78	.054		
	Greenhouse-Geisser	4.199	76.331	.055		
	Huynh-Feldt	4.199	78.000	.054		
	Lower-bound	4.199	39.000	.108		

Tests of Within-Subjects Contrasts Measure: Point-biserial Correlation

Source	Learning	Type III Sum of Squares	df	Mean Square	F	Sig.
Learning	Linear	.233	1	.233	4.016	.052
	Quadratic	.260	1	.260	5.247	.027
Error (Learning)	Linear	2.265	39	.058		
	Quadratic	1.934	39	.050		

Table 5. Repeated measures model continuation for the 3 exams indicating learning of self-assessment techniques. Pattern is shown in Figure 2.

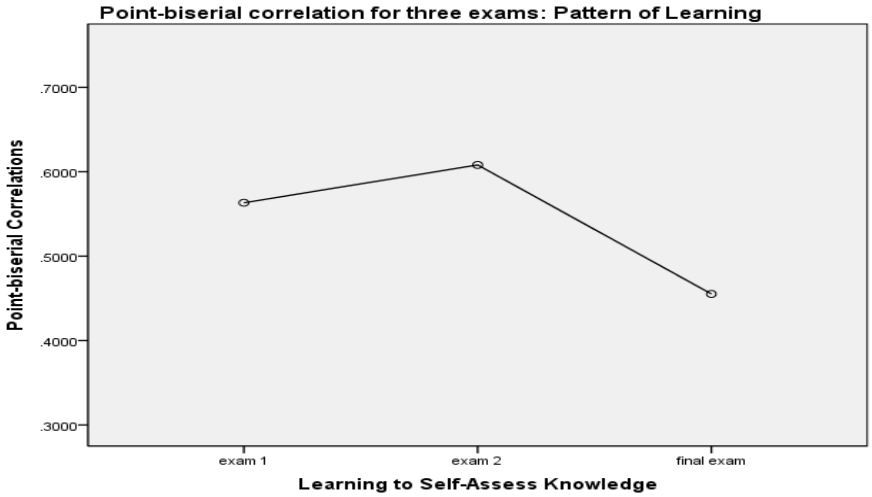

Figure 2. Pattern of learning appears quadratic when using the point-biserial correlation as a measure of self-assessment between correct response and self-perception of ability or problem difficulty.

% Acquisition indicating learning of exam taking abilities

Measure: % Acquisition **Within-Subjects Factors**

Learning	Dependent Variable
exam1	SCORE Exam 2
exam2	SCORE Exam 3
final	SCORE Final Exam

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Learning	Pillai's Trace	.668	42.287[a]	2.000	42.000	.000
	Wilks' Lambda	.332	42.287[a]	2.000	42.000	.000
	Hotelling Trace	2.014	42.287[a]	2.000	42.000	.000
	Roy's Largest	2.014	42.287[a]	2.000	42.000	.000

a. Exact statistic b. Design: Intercept Within Subjects Design: Learning

Table 6. Repeated measures model for the 3 exams indicates learning to take instructor's exams or something more permanent. See Figure 6 for pattern of learning.

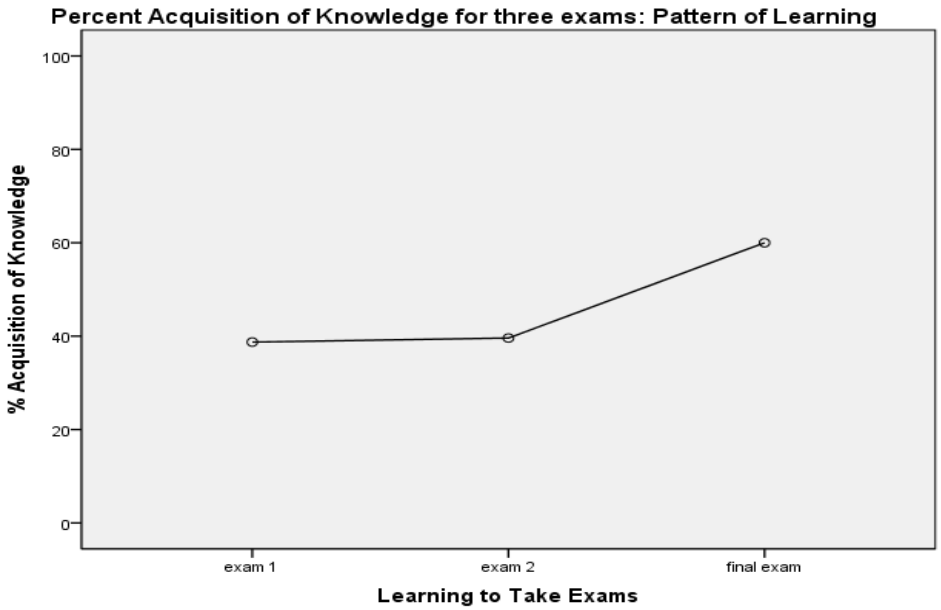

Figure 3. Pattern of learning appears quadratic in the reverse direction when using the exam score or % Acquisition of knowledge.

4. Summary

If educators are considering the self-study model of asking students how sure they are of the knowledge that they have obtained, at least in this setting of assigned surety to individual problems, we found that there is a similar pattern and association between correct responses and student confidence in a particular answer last year and this year in an introductory psychology courses. We found that a 10 point Likert scale worked better than a 5 point Likert scale. Additionally, students seemed to improve their ability to assess their confidence in answers on an exam very slightly when measured by the point-biserial correlation.

References

1. Dietz, Zachariah, Lovell, J. D., and Norton, J. A. (2011) " Including Student Ability to Assess Learning with Other Assessment Tools", *American Statistical Association 2011 Proceedings of the Section on Statistical Education*
2. Dietz, Zachariah, Lovell, J. D., Norton, Julia A. and Norton, John A (2012) "Student Self-Evaluation of Learning with Other Assessment Tools, Part II", *American Statistical Association 2012 Proceedings of the Section on Statistical Education*
3. Lovell, J. D. and Norton, J. A. (2002) "Percent Content Mastery Testing Of Content In College Courses", *Proceedings of the American Statistical Association Section on Statistical Education.*
4. Norton, J. A. and Lovell, J. D. (2000) "Repeated Measures Design In Assessment: Added Value of Instruction", *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Social Science*, pp. 282-283..

5. Norton, J. A., Zhou, Y., and Ganjeizadeh, F. (2008). “Better Features in Teaching Introductory Statistics”, *American Statistical Association 2008 Proceedings of the Section on Statistical Education.*
6. Norton, J. A. (2001) "Assessment of Introductory Statistics: Phase II", *Proceedings of the American Statistical Association Section on Statistical Education.*
7. Norton, J., J. Lovell, E. Suess, B. Trumbo, C. Sugahara, W. Rodriguez, J. Fowler, E. Worth, R. Young, T. Grube, V. Sue. (1999). "Statistics Program Assessment," *American Statistical Association 1999 Proceedings of the Section on Statistical Education.* 298-303.