Initial External Validation of REGRESS in Public Health Graduate Students

Kelley M. Kidwell, Ph.D.[1], and Felicity B. Enders, Ph.D., M.P.H.[2]

Abstract

Linear regression is typically taught as a second and potentially last required (bio)statistics course for Public Health and Clinical and Translational Science students. There has been much research on the attitudes of students toward basic biostatistics, but there has not been much assessing students' understanding of critical regression topics. The REGRESS (REsearch on Global Regression Expectations in StatisticS) quiz developed at Mayo Clinic utilizes 27 questions to assess understanding for simple and multiple linear regression. We performed an initial external validation of this tool with 117 University of Michigan public health students. We compare the results of pre- and postcourse quiz scores from the Michigan cohort to scores of Mayo medical students and professional statisticians. University of Michigan students performed higher than Mayo students on the precourse quiz due to previous related coursework, but did not perform as high postcourse indicating the need for course modification. In the Michigan cohort, REGRESS scores improved by a mean (standard deviation) of 4.6 (3.4), $p < 0.0001$. Our results support the use of the REGRESS quiz as a learning tool for students and an evaluation tool to identify topics for curricular improvement for teachers, while we highlight future directions of research. Clin Trans Sci 2014; Volume 7: 447–455

Keywords: linear regression, graduate level statistics course, biostatistics, assessment, validation

Introduction

Understanding statistics is becoming increasingly important in a data-saturated world. Technology growth has allowed the collection and analysis of huge data, but many students lack the skills to interpret and find relevance. This big data issue prompted the McKinsey Institute to release a May 2011 report explaining that "a significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning."[1] Demand could surpass supply by more than 140,000 in deep analytical positions by 2018 causing employers to request that colleges place more emphasis on critical thinking and complex reasoning.[2]

Statistics courses teach and instill these traits in students to help them become successful professionals. 2013 was named the International Year of Statistics to promote the practice of statistics and arouse interest in the field. A *New York Times* article from 2009 entitled "For Today's Graduate, Just One Word: Statistics" clarified that "statisticians are only a small part of an army of experts using modern statistical techniques for data analysis. Computing and numerical skills, experts say, matter far more than degrees."[3] A survey of 49 statistics instructors in the UK demonstrated that both the most important and difficult concepts in statistics courses include interpretation and relating results to the real world.[4]

Teaching analytical skills can be challenging, especially when the course is required and not elective and hence includes a broad mix of student expertise. However, these courses are exceedingly important for students to gain statistical fluency in order to communicate, understand, and tackle research problems efficiently. Particularly in fields of Public Health and Clinical and Translational Science, researchers from many backgrounds collaborate, often requiring basic statistical knowledge to set up and carry out investigations. At the University of Michigan, we attempt to arm graduate students in the School of Public Health with such statistical tools by requiring them to take at least one introductory statistical course.

Much of the assessment tools published in statistics education predict performance based on attitudes, anxiety, previous courses, gender, etc.,[5–16] but do not assess the efficacy of learned material or clarity of concepts. There are a few manuscripts examining the effectiveness of statistics courses to verify that students with broad backgrounds leave their programs understanding the fundamentals. One such piece developed and validated a 40 question outcome assessment survey for the typical undergraduate first course in statistics.[17] This assessment is widely used for intervention studies in statistics courses to refine course material and teaching techniques. Another assessed medicine residents' understanding of biostatistical concepts outside of a specific classroom context, focusing on their understanding of research literature in a survey including questions regarding demographics, statistics attitude, confidence of interpreting and assessing statistical concepts, and statistical knowledge.[18] Enders (2011)[19] discussed the previous two articles, as well as a few others, and concluded that a new tool for skills assessment for graduate level biostatistics was needed to aid statistics coursework and program development to improve graduate regression knowledge.

Such an assessment specific to linear regression was not previously available, but this material is often taught to a wide variety of students as the second course for undergraduate or graduate level studies. Enders (2013)[20] developed REGRESS (REsearch on Global Regression Expectations in StatisticS) to fill this gap. REGRESS is a competency-based construct built on CONSORT manuscript requirements (guidelines required by many medical and public health journals).[21,22] The quiz consists of 27 items intending to evaluate students' abilities to interpret and use the regression equation, understand modeling, statistical significance, and effect modification, and to assess assumptions, confounding, and colinearity. The REGRESS assessment is a practical and valuable tool for the students to assess their baseline knowledge and how much they have learned throughout the semester, while focusing on the most essential topics and analytic skills which they will need after the course is over. Moreover,

[1]Department of Biostatistics, University of Michigan, School of Public Health, Ann Arbor, Michigan, USA; [2]Mayo Clinic, Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Rochester, Minnesota, USA.

Correspondence: Kelley M. Kidwell (Kidwell@umich.edu)

DOI: 10.1111/cts.12190

the assessment can be useful for teachers as they can see how effective their teaching is to produce students who are proficient in understanding regression.

This manuscript presents the initial external validation of REGRESS for graduate students in public health completing a course on regression outside of Enders' classroom to help to verify that students from broad backgrounds understand the fundamental concepts of regression and interpretation. First we briefly explain the REGRESS instrument and present our statistical methods for validation and analysis. We then present results from the University of Michigan and compare these to Mayo Clinic. Finally, we discuss the apparent differences in results and provide suggestions on how the tool may be changed in future versions, as well as how to implement this tool both to assess students' knowledge before and after the course and to modify course plans to boost understanding.

Methods

REGRESS instrument

We provide a brief description here; for more details, please see Enders (2013).[20] The REGRESS instrument was developed for students and scientists in Clinical and Translational Science and Public Health to assess the major components of a regression course and those most useful for future research. We used version 5.1 of the instrument, as was used in Enders (2013).[20] The instrument consists of 27 questions organized into six domains: Interpreting and Using the Regression Equation (split into two domains, one for simple linear regression, SLR, and one for multiple linear regression, MLR); Modeling and Statistical Significance; Assessing Assumptions; Confounding and Colinearity; and Interaction. The quiz is available through an online link with an expected completion time of 45 minutes.

The first domain assesses understanding of SLR, in particular interpreting regression coefficients and understanding that association does not imply causation and results from the regression equation should only be used within the bounds of the data. Similarly, the second domain extends this to MLR seeking understanding of results from equations adjusting for other variables and linking graphs to equations. Statistical significance in single models and across nested models in SLR and MLR is tested through *p* values and confidence intervals. Assessing assumptions tests if students can find violations of linear regression in SLR and MLR. The next domain included both confounding and colinearity, which are more subtle issues in linear regression that may go unnoticed but are critically important in valid analyses. Especially in the context of observational studies, understanding confounding is integral for performing, reading, and interpreting public health research.[21–25] Without the understanding of confounding, incorrect conclusions and spurious associations are much more likely to be performed in analysis and missed under review. This domain assesses the students' abilities to identify scenarios under which these are likely to occur or have occurred. Finally, interaction is tested by means of matching equations and figures and understanding the interpretation of such an interaction with categorical or continuous predictors. This domain requires students to understand how each type of predictor variable acts within an MLR model (i.e., a model with a binary and a continuous predictor will make two lines) and how interaction changes the graphical interpretation of the model. The added complexity within this domain was included to permit assessment of interaction through visual means rather than requiring complex calculations during the quiz.

We have not included the instrument so that students will not be exposed to it prior to use in the classroom, but a copy of the instrument is available from Felicity Enders (Enders.Felicity@mayo.edu).

University of Michigan and Mayo cohorts

After creating the REGRESS instrument, Enders internally validated the tool using graduate students enrolled in her classes through the Clinical and Translational Science program at Mayo Clinic. The Mayo student cohort spanned students enrolled in Enders' classes from 2010 to 2012, where the class covered topics associated with the domains of the quiz, as well as an introduction to logistic and Cox regression. Fifty-two students took the quiz precourse and 59 students completed the quiz postcourse. The subset of students who completed both the pre- and postcourse assessments was small, so that in this previous analysis, independent sample tests were used and considered conservative. The Mayo student cohort consisted of academic clinicians of whom 47% were males who had never studied linear regression (89%). A convenience sample of 22 practicing statisticians at the Mayo Clinic with an MS or PhD in statistics or a related field were also asked to complete the survey to compare results between students and statisticians in the field. Selected data republished with permission.

As an initial external validation, the REGRESS instrument was implemented at the University of Michigan (UM) as part of a linear regression course offered to public health students during the winter semesters of 2013 and 2014. This course is the second biostatistics course offered and generally the last course taken for most public health majors (besides those enrolled in a biostatistics or epidemiology degree program). The course focuses on linear regression, both SLR and MLR, dedicating half of the semester to this topic with the other half dedicated to logistic, Poisson, and Cox regression. Concepts such as interpreting SLR and MLR, understanding interaction, confounding, and colinearity are repeated throughout the semester. The instructor completed her PhD in biostatistics in 2012 and 2013 was her first time teaching this course at UM. She based her course materials upon those of the previous instructor, who had taught the course the previous 4 years to a similar student base.

For the 2014 semester, lectures were modified by including three to six questions during each lecture where the students used clickers to choose the correct multiple choice answer. These questions included past exam problems and questions from online sources[26–31] and were proposed to engage the class. Specifically, the questions challenged students' understanding of the main concepts learned in class that day and focused on the enforcement of understanding plots and how these relate to equations and interpretations of SLR and MLR. The 2014 cohort was also asked for qualitative understanding after each REGRESS question to qualitatively assess confusion with the questions. We received IRB approval to offer the pre- and postcourse REGRESS quiz where students could opt out of being included in the research aspect. A total of 128 students were enrolled in the course in 2013 and 2014, but 11 students opted out of the research component, leaving a final sample of 117 students who completed the pre- and postcourse quiz.

The same UM students completed both the precourse (within the first 2 weeks of the semester) and postcourse surveys. In

2013, students had the option to take the postcourse quiz upon completing the linear regression material (half-way through the semester) or at the end of the semester, but all students completed the postcourse survey directly following linear regression material in 2014. It was expected that students could perform well after only the linear regression related material. The time between pre- and postcourse REGRESS completion ranged from 38 to 103 days with a median time to completion of 58 days from the precourse quiz. We refer to the postcourse REGRESS for all 117 of these students, regardless if they completed the quiz during or after the semester.

Statistical analysis

Nonresponse or response of "I don't know" was coded as incorrect. The overall REGRESS score (questions 1–27) and SLR score (questions 1–11) are presented with the mean and standard deviation for each group at each time. For the UM cohort, pre- and postcourse overall, SLR, and domain data were compared using paired *t*-tests or sign tests. Individual items for the UM cohort were assessed using the sign test on pre- and postcourse score differences. Mayo cohort values are taken from Enders 2013 where methods are given.[19]

Comparison of summary scores between the cohorts used *t*-tests ignoring multiple comparisons corrections. Comparison of the frequency of correct answers of individual items between cohorts used Fisher's exact test. Internal reliability was calculated for each score using Cronbach's alpha using pre- and postcourse responses from the UM cohort. We also performed a sensitivity analysis for the difference in the pre- and postcourse scores for UM students from the 2013 and 2014 cohorts with a *t*-test. The change in scores over time was compared between cohorts with a Wilcoxon rank-sum test. To investigate the relationship of REGRESS scores with course grades, we separated course grades into those which were linear regression (LR) related and those which were post-LR material (logistic, Poisson, and Cox regression). The post-LR concepts repeated several of those from LR (confounding, interaction, colinearity), so we were interested if success on REGRESS was correlated to success outside of linear regression. These relationships between pre- and postcourse scores to LR related and post-LR related average course grades for UM students was assessed using spearman correlation. LR related average grades included four homeworks (worth 40–48 points each), a midterm (100 points), and an article critique (40 points). The post-LR-related average grades included 2 homeworks (each worth 47 points), a quiz (54 points), and a final exam (97 points). These related course grades were calculated as a weighted average to reflect the weighting used for the overall course grade. Finally, a linear regression on the change in REGRESS scores (postcourse scores minus precourse scores) with predictors of precourse score, gender, year (2013 or 2014), and time in days between the pre- and postcourse quiz assessed associations between these variables and the change in scores. Two-sided 5% type I errors were used throughout. The SAS statistical software package was used (version 9.3, Cary, NC, USA).

Results

Table 1 presents the characteristics of these students. Primarily, the UM cohort consists of masters level students (90%, mostly MPH, 9 students with MS degrees) who had taken a statistics course where regression was one of the topics in the past year (81%). There were more females in the course (73%) with 57% of the students seeking a master's in environmental health science, 23% in health behavior and health economics, 14% in health management and policy, 2% in epidemiology, and 4% in disciplines outside of public health. Only two people had an undergraduate degree in statistics or a related field and had completed it within the last 2 years. This cohort is similar in characteristics to all graduate students in accredited public health schools (46 schools in 2010) where most graduate students are female (71%) earning master's degrees (86%, ASPH 2010 Annual Data Report),[32] but does differ from the Mayo cohort as Mayo consisted of Clinical and Translational Science students with more males (47%) without previous coursework in statistics.[20]

Figure 1 and *Table 2* present results for the overall REGRESS score in the UM cohort. Total scores from the precourse quiz ranged from 4 to 21 (out of 27), with a mean score of 11.6 (SD = 3.1). Total scores from the postcourse quiz ranged from 8 to 25 with a mean score of 16.2 (SD = 3.3). Eight students (6.8%) obtained a lower score on the postcourse quiz than on the precourse quiz, whereas the other 109 (93.2%) performed better by 1 to 16 points. There was with a statistically significant increase in postcourse score compared to precourse score with a mean of 4.6 (SD = 3.4), $p < 0.0001$.

The 2013 and 2014 cohorts had a similar precourse REGRESS score (mean [SD] of 11.3 [2.8] for 2013 and 11.8 [3.4] for 2014; $p = 0.35$). However, postcourse REGRESS scores were slightly higher for the 2014 cohort (mean [SD] of 15.2 [3.1] for 2013 and 17.0 [3.3] for 2014; $p = 0.002$). The change in REGRESS scores also showed a slight increase for 2014 (mean [SD] of 3.9 [3.0] for 2013 and 5.2 [3.6] for 2014; $p = 0.027$). Although statistically significant, the differences in scores were small, so the cohorts are consequently presented as one group for the remainder of the analysis.

Table 2 presents the average SLR and domain scores for the UM cohort. The mean precourse SLR score was 7.4 (SD = 1.6) and postcourse score was 8.4 (SD = 1.8). This average increase of 1 point for all students is statistically significant, $p < 0.0001$. Unfortunately, however, 25 (21%) students performed lower on the postcourse SLR quiz by 1 to 4 points, 22 (19%) obtained the same score, but 70 (60%) students performed better by 1 to 16 points. Each domain increased from precourse to postcourse by small but statistically significant amounts. Most domains increased by a median value of 1 point, which we consider a meaningful increase from pre- to postcourse scores due to the small number of questions included in each domain, but the median change for the domains of assessing assumptions and interaction was zero due to many incorrect answers at both time points.

Table 3 compares the mean pre- and postcourse scores for the UM cohort as compared to the Mayo students and statisticians. UM students performed significantly higher on the precourse quiz than Mayo students ($p = 0.001$). This increased level of knowledge is due to the majority of students having taken a course just prior to enrolling in the regression course which introduces SLR at the end of the course. The UM students, however, performed significantly lower postcourse when compared to either the Mayo student cohort or statisticians ($p < 0.0001$).

To understand the performance of UM students from pre- to postcourse and to further compare the UM and Mayo cohorts, we examined the distribution of correct responses to the individual items. UM students significantly improved in 16 items (59%) from pre- to postcourse quiz. These items include three questions from domain 1: SLR (items 4, 6, and 7, change in mean outcome, predicting beyond data, and interpreting unadjusted slope), two

Characteristic	N	%
Gender		
Male	32	27.4
Female	85	72.6
Degree		
Master's (MPH, MS)	105	89.7
MD/MPH	3	2.6
PhD*	9	7.7
Master Level Programs ($n = 108$)		
Environmental Health Science	62	57.4
Health Behavior and Health Economics	25	23.1
Health Management and Policy	15	13.9
Epidemiology	2	1.9
Other (outside Public Health)	4	3.7
Extent Studied Linear Regression		
Never	2	1.7
Stat course where regression was one of topics	112	95.7
One course primarily focused on regression	2	1.7
Two or more courses on regression	1	0.9
When was the last time you took a course with regression		
Never	2	1.7
Currently enrolled in first class	2	1.7
Ended <12 months ago	95	81.2
Ended 12–24 months ago	11	9.4
Ended >24 months	7	6.0
Current level of expertise for Regression		
Never use	3	2.6
Have used only for homework and activities during class	100	85.5
Used outside class sporadically, not during past 24 months	6	5.1
Used outside class often, at least once in past 24 months	6	5.1
Use outside class regularly, more than twice in past 24 months	2	1.7
Degree in Statistics or Related Field		
No	115	98.3
Undergraduate degree	2	1.7
If degree, how long ago completed?		
No statistics degree	115	98.3
0–2 years	2	1.7

*PhDs were in Anthropology, Toxicology, Epidemiology, Nursing, Molecular, Cellular and Developmental Biology, Kinesiology, Psychology.

Table 1. Characteristics of UM student samples from 2013 and 2014.

from domain 2: MLR (items 14 and 22, linking graphs to equations and interpreting adjusted slope), three in modeling and statistical significance (items 5, 23, and 26, how to test association and selecting nested models), two from each assessing assumptions (items 17 and 18, assessing homoscedasticity and normality of errors), three from confounding and colinearity (items 19, 27, and 20 predicting and diagnosing confounding and predicting colinearity), and all three items in interaction (items 13, 15, and 25, linking graphs to equations and diagnosing interaction). The improved scores for assessing assumptions (besides homoscedasticity), confounding and colinearity, and interaction were statistically significant, but the proportion of correct answers remained low at 9–62%. The UM cohort performed slightly worse postcourse on two items assessing assumptions (11, presence of outliers and 16, assessing independence). The eight other items improved, but not significantly.

Comparing the UM cohort to Mayo students, we see from *Table 4* that UM students performed statistically significantly higher than Mayo students precourse on items 4 (assessing slope), 5 (test association), 10 (correlation), 12 (linking graph to equation with continuous and categorical predictors), 21 (predicting outcome), and 25 (diagnosing interaction). For the postcourse quiz, UM students scored significantly higher for items 6 (predicting beyond data) and 10 (assessing strength of correlation), but significantly lower for 12 (44%) items. The most notable differences where UM students scored much lower than the Mayo cohort most often involved inference from graphs (items 2, 3, 13, 14, 15, 17, 18, 19). The UM cohort also struggled with four other items (16, assessing independence; 20, predicting colinearity; 25, diagnosing interaction; 27, diagnosing confounding), not able to match the response of Mayo students postcourse.

Comparing the UM student postcourse scores to statisticians, statisticians scored significantly higher for 7 (26%) of the items (specifically five graphical items 2, 3, 14, 15, and 19; and two other items including 16, assessing independence; and 25, diagnosing interaction). UM students performed similarly to the statisticians on the remaining 20 questions. Interestingly, 36% or more of the statisticians answered multiple graphing questions incorrectly: items 2 (finding the y-intercept), 12, 13, and 14 (linking equations to graphs), 18 (assessing normality of errors), and 20 (predicting colinearity), so that the lack of statistical difference from UM students does not reflect high scores from both groups.

Internal reliability of the quiz was assessed using all questions both pre- and postcourse for the UM cohort as shown in *Table 5*. Cronbach's alpha for the overall REGRESS score was 0.7 compared to 0.9 for all Mayo cohorts at all times. For the SLR score, the Cronbach's alpha was 0.6 compared to 0.7 for the Mayo cohorts. For the domain scores, Cronbach's alpha ranged from 0 to 0.6 compared to 0.3 to 0.8 for the Mayo cohort.

Furthermore, we were interested in evaluating the correlation between REGRESS score and course performance in the UM

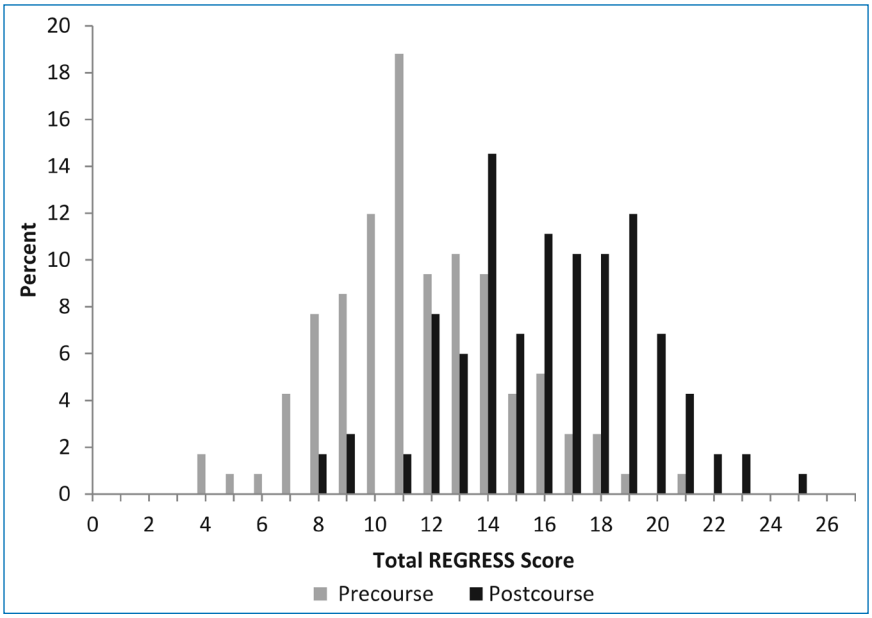

Figure 1. Distribution of REGRESS scores by pre- and postcourse.

cohort in *Figure 2*. We expect a valid instrument to have high correlation with the assessment of course material conditioning on the class teaching critical regression concepts. Pre- and postcourse REGRESS scores were positively related with Spearman correlation coefficient of 0.5, $p < 0.001$. This relationship may reflect those with a higher affinity for statistics, as well as those who took the quiz more seriously at both time points. Precourse scores were moderately correlated to the overall course grade ($r = 0.3, p < 0.001$) and slightly correlated to average grades throughout the semester split by those based on LR related material ($r = 0.2$, $p = 0.05$) and post-LR related material ($r = 0.2$, $p = 0.06$). The postcourse REGRESS score was significantly and more strongly correlated to the overall postcourse average grade ($r = 0.5$, $p < 0.001$). When split by subject matter, the postcourse REGRESS score was significantly associated with both the LR related average grades ($r = 0.5, p < 0.001$) and the post-LR related average grades ($r = 0.4, p < 0.001$) with higher REGRESS scores associated with higher LR or post-LR related grades.

Results from linear regression (*Table 6*) reveal that the precourse score and year were significantly related to the change in REGRESS score ($p < 0.001$ and $p = 0.007$, respectively), but sex and time in days between taking the pre- and postcourse quiz were not statistically significant ($p = 0.24$ and $p = 0.30$, respectively). The higher the precourse REGRESS score, the less change from precourse to postcourse score, exhibiting a ceiling effect. 2014 students performed better on the postcourse quiz and therefore had a larger difference in their scores (on average 2.2 points, $p = 0.007$). These variables, however, only explained 30% of the variation in postcourse REGRESS scores.

Discussion

The REGRESS quiz is a tool to assess students' understanding of the important topics in linear regression. Although the UM students significantly improved scores from the precourse to postcourse quiz ($p < 0.0001$), there is still more to be desired in postcourse quiz performance. The UM cohort had lower postcourse quiz scores than the internal Mayo students ($p < 0.0001$) and practicing statisticians ($p < 0.0001$); in our external setting the average postcourse REGRESS score was 16.2 of 27 points. We also observed a small but statistically significant increase for all six REGRESS domain scores ($p < 0.0001$).

The UM students mainly struggled with interpreting questions and concepts from graphical displays of data (items 2, 3, 13, 14, 15, 18, 19). Of these items, while most did significantly improve from pre- to postcourse, the correct response rates remained low post course.

The poor performance on items 2 and 3 is not surprising as these two questions are related, so those who answered item 2 incorrectly were also likely to answer item 3 incorrectly. Item 2 requires a careful reading and understanding of the axes limits. Those who completed the quiz quickly were likely to miss that the graph origin is not (0,0). In the UM cohort, 55% chose the answer option of 0 which would be correct if the origin was (0,0) and correspondingly chose the incorrect solution for item 3. Regardless of the potentially "tricky" question, students should be trained to carefully examine plots and to be able to relate graphical results to equations and to sentences explaining the results.

The 2014 UM cohort had a marginal but statistically significant increase in postcourse scores and change in scores than the 2013 cohort, despite a concerted effort to add clicker questions focusing on interpreting and matching equations to graphs. The qualitative data provided by students on the graph questions suggested that their knowledge on this topic was still not robust. For example, one student wrote, "I still find these kinds of variable and equation/graph interpretations hard as I have difficulty determining what the different types of variables would look like graphed out." Another student wrote, "I remember the clicker question that was similar, but I also struggled with that one." Together, these data suggest that this version of the REGRESS is sensitive enough to capture changes in knowledge attributable to an intervention, but that it is challenging to provide a deep enough understanding of difficult topics for students to transfer that knowledge to unfamiliar question formats.

Unfortunately the 2013 students generally learned their weaknesses at the end of the course and were unlikely to review material to clarify concepts. To utilize the quiz to its full potential as a learning tool, the 2014 group took the REGRESS right after linear regression material and had a review on REGRESS material in class to clarify concepts which were not highly scored on the postcourse quiz. By giving the students their REGRESS score upon completion, the students were able to see how much they learned throughout the semester, as well as identify topics which remained challenging for subsequent review.

Additionally, in both years, the REGRESS exercise allowed the professor to understand strengths and weaknesses in curriculum. For example, since the 2013 UM cohort did not perform well on the items related to graphs, the regression course at UM was modified to include more graphics and interactive questions regarding these graphics for interpretation. The 2014 cohort results suggest further modifications may still be needed for this topic.

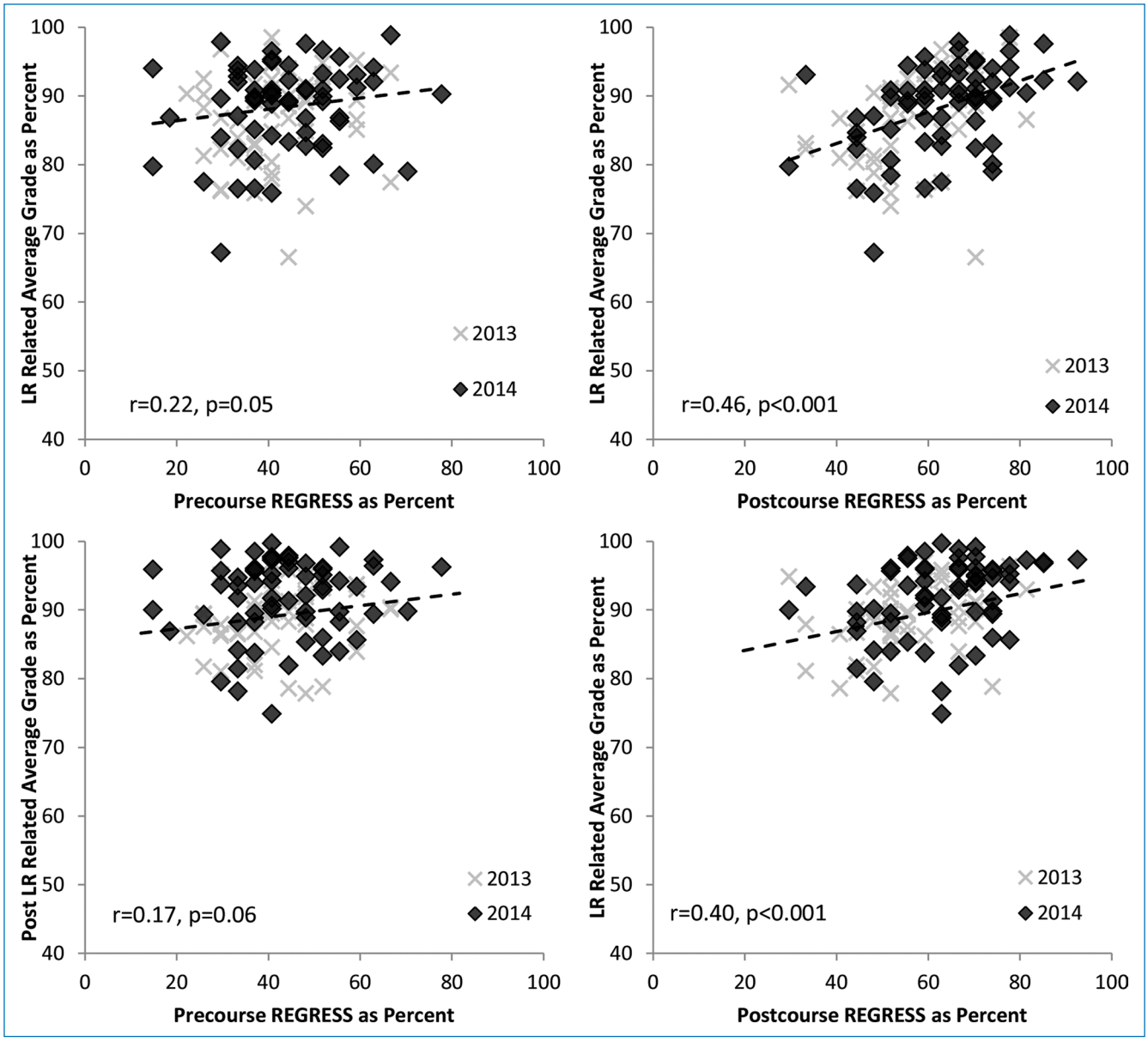

Figure 2. Relationship of pre- and postcourse REGRESS scores to linear regression related (LR related) and post-LR related average grades. LR related average grades include four homeworks, a midterm, and an article critique. The post-LR related average grades include 2 homeworks, a quiz, and a final exam. The Spearman correlation coefficient is denoted by *r* in the bottom-right corner with its corresponding *p* value assessing significant correlation.

Scores	Precourse	Postcourse	Change in Scores, Post – Pre	Pre vs. Post
Summary	Mean (SD)	Mean (SD)	Mean (SD)	*p* Value*
REGRESS (out of 27)	11.6 (3.1)	16.2 (3.3)	4.6 (3.4)	<0.0001
SLR Score (out of 11)	7.4 (1.6)	8.4 (1.8)	1.0 (1.8)	<0.0001
Domain	Median (range)	Median (range)	Median (Range)	
Interpreting and Using SLR Equation (of 8)	5.0 (1.0–8.0)	6.0 (1.0–8.0)	1.0 (–4.0–6.0)	<0.0001
Interpreting and using MLR Equation (of 4)	2.0 (0–4.0)	3.0 (0–4)	1.0 (–2.0–4.0)	<0.0001
Modeling and Statistical Significance (of 4)	2.0 (0–4.0)	3.0 (1–4)	1.0 (–2.0–3.0)	<0.0001
Assessing Assumptions (of 4)	1.0 (0–3.0)	2.0 (0–3.0)	0 (–2.0–3.0)	<0.0001
Confounding and Colinearity (of 4)	1.0 (0–4.0)	2.0 (0–4.0)	1.0 (–3.0–4.0)	<0.0001
Interaction (of 3)	1.0 (0–3.0)	1.0 (0–3.0)	0 (–2.0–3.0)	<0.0001

*p Value from paired *t*-test for summary scores and Wilcoxon signed rank test for domain scores.

Table 2. Summary of UM student scores.

REGRESS scores (out of 27)	N	Mean (SD)	Range	p Value†
Precourse				
UM students	117	11.6 (3.1)	4–21	
Mayo students	52	9.3 (4.3)	0–21	0.001
Postcourse				
UM students	117	16.2 (3.3)	8–25	
Mayo students	59	19.0 (3.5)	10–27	<0.0001
Statisticians	22	20.1 (3.5)	13–24	<0.0001

*Data from Mayo students and practicing statisticians reprinted with permission from Enders (2013).
†*p* Value from *t*-test comparing scrors from UM students to Mayo students and UM student to Mayo statisticians.
SD=standard deviation.

Table 3. Comparison of mean scores between UM students and Mayo results.*

UM students scored higher at baseline than Mayo students due to having more prior experience with regression within the context of a previous course. Ninety-six percent of all UM students had taken a prior course where regression was one of the topics covered (most students had taken an introduction to biostatistics course the previous semester which introduced regression and one-way ANOVA at the end of the semester). Consequently, for these students, the SLR score is likely not useful and we see no meaningful difference in pre- and postcourse scores. The SLR score may be more appropriate for those engaged in their first regression course, especially at the undergraduate level.

Differences between the UM students and Mayo students were not unexpected for several reasons. First, the REGRESS was developed by Dr. Enders, who also taught the Mayo students. Even though Dr. Enders did not teach with the goal of high REGRESS quiz performance, the same concepts were naturally emphasized since she was the creator of the test. Another major difference was due to the structure of the courses and difference in instructors between UM and Mayo. The course content was similar, but the time spent on each topic and mode of presentation varied. Dr. Enders also has more experience teaching this subject matter since she graduated in 2003 and has been a primary course instructor at the graduate level since 2000, whereas the 2013 course was part of Dr. Kidwell's first year teaching after graduating in 2012. Furthermore, the makeup of the students differed. The Mayo students were generally physicians or PhD students engaged in medical research, whereas the UM students were generally master's-level Public Health students who may or may not continue to careers with a research component. While both cohorts mainly consisted of students who were required to take the course, their attitudes toward the course and motivation within the course most likely varied. A limitation to our study is the lack of qualitative assessments between cohorts.

The smaller levels of Cronbach's alpha from the UM cohort were likely due to several reasons. First, one expects higher internal reliability from Mayo since the creator of the REGRESS instrument was also the course professor, whereas at UM, the REGRESS test was not taken into account in course planning in 2013. While the course was modified in 2014 to emphasize graphs by adding clicker questions, general emphasis within lectures remained constant. Additionally, the lower levels of Cronbach's alpha using the UM cohort, especially for assessing assumptions, were due to two questions in the domain with very low scores; students scored well on the other questions in the domain. With continued use of the REGRESS tool with a variety of students, we can better estimate the internal reliability of the quiz. Additional studies will be needed to further validate the REGRESS quiz outside of Mayo, as different student populations, instructors, course material, and teaching style may impact difference in pre- versus postcourse scores and further influence modifications for future versions of the REGRESS instrument.

Supporting internal validity of the quiz, the postcourse REGRESS score was significantly correlated with LR related and post-LR related average scores for those in the UM cohort. The material learned from linear regression relates to that seen in logistic, Poisson, and Cox regression, so we expect a correlation between the postcourse REGRESS score and post-LR related material, but to be less than that of the LR related material, which was seen.

From our results using the REGRESS as a teaching evaluation, assessing student performance and internal validity in an external sample, we hypothesize that either minor modifications to REGRESS are necessary or that specifics of the external cohort led to lower performance. Further modification to the University of Michigan course will be made to continue to emphasize interpretation of graphs.

These results will also help guide the next version of the REGRESS. Due to the qualitative assessments from the 2014 cohort about confusion of questions in REGRESS, we intend to make minor modifications to the REGRESS quiz to clarify questions. Moreover, modifications to the domains of interaction and assessing assumptions may be particularly important. These domains may either contain concepts which differ too broadly or require a greater number of items to represent the domain.

The strengths of this study include the use of a sample of students not influenced by the creator of the REGRESS tool. We were also able to ensure all students at UM took both the precourse and postcourse REGRESS quiz. Our sample size was also robust to capture not only overall changes in score from pre- to postcourse but also discriminate between the 2013 and 2014 UM cohorts and assess scores within each domain. However, our assessment also has weaknesses. For explicit external validation of the tool, more data from different classrooms are necessary, as well as additional qualitative item assessments. We have also combined the 2013 and 2014 UM cohorts for the majority of our analyses.

As a result of this work, we will refine the REGRESS quiz. We will continue to implement the modified quiz in regression courses to validate the new version and consider further potential modifications to REGRESS to better capture students'

Item number and description	UM students			Mayo students				Mayo professional statisticians	
	pre-course *N* (%)	post-course *N* (%)	*p* Value pre vs. post	pre-course *N* (%)	*p* Value pre‡	post-course *N* (%)	*p* Value post§	*N* (%)	*p* Value post**
Interpreting & using SLR equation									
1 Find slope from graph†	109 (93)	108 (92)	1.00	44 (85)	0.09	56 (95)	0.75	21 (96)	1.00
2 Find y-intercept from graph†	24 (21)	35 (30)	0.06	8 (15)	0.53	29 (49)	0.020	14 (64)	0.004
3 Link graph to equation†	27 (23)	35 (30)	0.23	18 (35)	0.13	28 (48)	0.030	16 (73)	<0.001
4 Change in mean Y given X†	83 (71)	104 (89)	0.001	22 (42)	0.001	57 (97)	0.09	19 (86)	0.72
6 Predicting beyond data†	64 (55)	100 (85)	<0.001	31 (60)	0.62	25 (42)	<0.001	18 (82)	0.75
7 Interpret slope, unadjusted†	84 (72)	102 (87)	0.004	32 (62)	0.21	57 (97)	0.06	21 (96)	0.47
8 Association is not causation†	81 (69)	85 (73)	0.56	38 (73)	0.72	51 (86)	0.06	20 (91)	0.10
10 Assess strength of correlation†	95 (81)	103 (88)	0.10	30 (58)	0.002	37 (63)	<0.001	22 (100)	0.13
Interpreting & using MLR equation									
12 Link graph to equation, continuous and categorical Xs	57 (49)	67 (57)	0.23	13 (25)	0.004	32 (54)	0.75	14 (64)	0.64
14 Link graph to equation, >1 continuous and binary Xs	13 (11)	38 (32)	<0.001	10 (19)	0.22	36 (61)	<0.001	14 (64)	0.008
21 Predict Y, categorical X	97 (83)	107 (91)	0.08	24 (46)	<0.001	52 (88)	0.59	22 (100)	0.36
22 Interpret slope, adjusted	45 (38)	96 (82)	<0.001	14 (28)	0.16	51 (91)	0.52	20 (91)	0.53
Modeling & statistical significance									
5 How to test association†	100 (85)	115 (98)	0.001	26 (50)	<0.001	58 (98)	1.00	22 (100)	1.00
23 Select nested model, not statistically significant	7 (6)	53 (45)	<0.001	3 (6)	1.00	30 (51)	0.52	11 (50)	0.82
26 Select nested model, statistically significant	38 (32)	90 (77)	<0.001	11 (21)	0.15	49 (83)	0.43	16 (73)	0.79
9 Relationship between sample size and statistical significance†	107 (91)	109 (93)	0.80	45 (87)	0.41	54 (92)	0.76	22 (100)	0.36
Assessing assumptions									
11 Presence of outliers†	91 (78)	86 (74)	0.44	35 (67)	0.18	39 (66)	0.38	20 (91)	0.10
16 Assess independence	16 (14)	7 (6)	0.049	12 (23)	0.18	13 (22)	0.003	5 (23)	0.024
17 Assess homoscedasticity	36 (31)	94 (80)	<0.001	13 (25)	0.47	57 (97)	0.003	20 (91)	0.37
18 Assess normality of error terms	2 (2)	10 (9)	0.022	5 (10)	0.029	30 (51)	<0.001	5 (23)	0.06
Confounding & colinearity									
19 Predict confounding	31 (26)	48 (41)	0.008	8 (15)	0.17	49 (83)	<0.001	16 (73)	0.009
27 Diagnose confounding	17 (15)	73 (62)	<0.001	6 (12)	0.81	49 (83)	0.006	13 (59)	0.81
20 Predict colinearity	27 (23)	57 (49)	<0.001	8 (15)	0.31	39 (66)	0.037	13 (59)	0.49
24 Diagnose colinearity	27 (23)	38 (32)	0.14	9 (17)	0.54	14 (24)	0.29	10 (46)	0.33
Interaction									
13 Link graph to equation, continuous and binary interaction	20 (17)	34 (29)	0.024	3 (6)	0.054	34 (58)	<0.001	12 (55)	0.027
15 Link graph to equation, continuous interaction	19 (16)	39 (33)	0.001	5 (10)	0.34	39 (66)	<0.001	19 (86)	<0.001
25 Diagnose interaction	39 (33)	63 (54)	0.001	9 (18)	0.042	55 (93)	<0.001	17 (77)	0.06

*Data from Mayo students and practicing statisticians reprinted with permission from Enders 2013.
†Denotes items included in the SLR score
‡*p* Value comparing UM students ($N = 117$) to Mayo students ($N = 52$), precourse scores
§*p* Value comparing UM students ($N = 117$) to Mayo students ($N = 59$), postcourse scores
**p Value comparing UM students postcourse scores ($N = 117$) to statisticians ($N = 22$)

Table 4. Table of the frequency and proportion of correct pre- and postcourse REGRESS answers for each individual question. Comparisons are made between the 52 UM student cohort, 52 students at Mayo who took the precourse quiz, 59 students at May who took the postcourse quiz and to 22 professional statisticians*.

Summary scores	Cronbach's alpha
REGRESS score (of 27)	0.7
SLR score (of 11)	0.6
Domain scores	
Interpreting & using SLR equation (of 8)	0.6
Interpreting & Using MLR equation (of 4)	0.3
Modeling & statistical significance (of 4)	0.3
Assessing assumptions (of 4)	0
Confounding & colinearity (of 4)	0.3
Interaction (of 3)	0.2

Table 5. Internal reliability of summary and domain scores based on the UM cohort pre- and postcourse REGRESS scores.

Variable	Coefficient	Standard error	*p* Value
Intercept	8.89	2.29	<0.001
Precourse score	−0.54	0.09	<0.001
Female	0.74	0.62	0.24
2014 vs. 2013	2.20	0.79	0.007
Time from precourse taken (days)	0.03	0.03	0.30

Table 6. Multiple linear regression on change in REGRESS score (postcourse minus precourse scores).

understanding of critical regression topics. We hope to include more institutions in future studies to better assess the REGRESS in different settings.

Conclusion

Using the UM results, we provide an initial external validation of the REGRESS tool as an assessment of student understanding and as an instrument to guide instructors in identifying topics for revision in course materials. The REGRESS quiz can be useful for students, instructors, and researchers. Students can use their REGRESS scores to identify topics for further study and assess their understanding of critical regression concepts. The REGRESS quiz is also useful for instructors to provide feedback on topics to consider updating in the future. The REGRESS quiz can also be used as an outcome measure to quantify the impact of educational interventions. . Our work demonstrates that the REGRESS quiz can be successfully used outside the original Mayo Clinic setting to measure students' understanding of critical concepts in linear regression.

Acknowledgments

This material is based upon work supported by the Center for Research on Learning and Teaching's Investigating Student Learning Grant.

References

1. Manyika J, Chui M, Bughin J, Brown B, Dobbs R, Roxburgh C, Byers AH. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. 2011

2. Hart Research Associates. 2010. *Raising the Bar: Employers' Views on College Learning in the Wake of the Economic Downturn*. Washington, DC: Association of American Colleges and Universities.

3. Lohr S. For Today's graduate, just one word: Statistics, August 5, 2009. *New York Times*. http://www.nytimes.com/2009/08/06/technology/06stats.html?_r = 0. Accessed October 3, 2013

4. Nicholson JR, Mulhern G. (2000). Conceptual challenges facing A level statistics students: Teacher and examiner perspectives. Papers on Teaching and Learning Statistics –ICME9. http://iase-web.org/documents/papers/icme9/ICME9_03.pdf. Assessed February 26, 2014.

5. Elmore P, Vasu ES. Relationship between selected variables and statistics achievement: building a theoretical model. *J Educ Psychol.* 1980; 72(4): 457–467.

6. Elmore P, Vasu ES. A model of statistics achievement using spatial ability, feminist attitudes and mathematics-related variables as predictors. *Educ Psychol Meas*. 1986; 46(1): 215–222.

7. Schau C, Stevens J, Dauphinee TL, Del Vecchio A. The development and validation of the survey of attitudes toward statistics. *Educ Psychol Meas.* 1995; 55: 868–875.

8. Schram CM. A meta-analysis of gender differences in applied statistics achievement. *J Educ Behav Stat.* 1996; 21(1): 55–70.

9. Tremblay PF, Gardner RC, Heipel G. A model of the relationships among measures of affect, aptitude, and performance in introductory statistics, *Can J Behav Sci.* 2000; 32(1): 40–48.

10. Baloglu M. Individual differences in statistics anxiety among college students. *Personal Indiv Differ.* 2003; 34(5): 855–865.

11. Bandalos DL, Finney SJ, Geske JA. A model of statistics performance based on achievement goal theory. *J Educ Psychol.* 2003; 95(3): 604–616.

12. Nasser FM. Structural model of the effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *J Stat Educ [Online]* 2004; 12(1), Retrieved October 3, 2013, from http://www.amstat.org/publications/jse/v12n1/nasser.html

13. Cashin SE, Elmore PB. The survey of attitudes toward statistics scale: a construct validity study. *Educ Psychol Meas.* 2005; 65(3): 509–524.

14. Tempelaar DT, Schim van der Loeff S, Gijselaers WH. A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities, and course performance. *Stat Educ Res J.* 2007; 6(2): 78–102.

15. Chiesi F, Primi C. Cognitive and non-cognitive factors related to students' statistics achievement. *Stat Educ Res J.* 2010; 9(1): 6–26.

16. Emmioğlu E, Capa-Aydin Y. Attitudes and achievement in statistics: a meta-analysis study. *Stat Educ Res J.* 2012; 11(2): 95–102.

17. delMas R, Garfield J, Ooms A, Chance B. Assessing students' conceptual understanding after a first course in statistics. *Stat Educ Res J.* 2007; 6(2): 28–58. http://www.stat.auckland.ac.nz/ iase/serj/SERJ6(2)_delMas.pdf

18. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *J Am Med Assoc.* 2007; 298(9): 1010–1022.

19. Enders F. Evaluating mastery of biostatistics for medical researchers: need for a new assessment tool. *Clin Trans Sci.* 2011; 4: 448–454.

20. Enders F. Do clinical and translational science graduate students understand linear regression? Development and early validation of the REGRESS quiz. *Clin Trans Sci.* 2013: doi: 10.1111/cts.12088

21. Armstrong R, Waters E, Moore L, Riggs E, Cuervo LG, Lumbiganon P, Hawe P. Improving the reporting of public health intervention research: advancing TREND and CONSORT. *J Public Health.* 2008; 30(1): 103–109.

22. Moher D, Hopewell S, Schulz KF, Montori V, Gotzche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010; 340(c869): 1–28.

23. Des Jarlais D, Lyles C, Crepaz N, Group T. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health.* 2004; 94(3): 361–366.

24. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ.* 2007; 335(7624): 806–808.

25. Vandenbroucke JP, von Elm E, Altman DG, Gotzche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med.* 2007; 4(10): 1628–1654.

26. Murphy TJ, McKnight C, Richman M, Terry R. Clicker Questions. 2009, July 29: Retrieved from http://www.ou.edu/statsclickers/clickerQuestions.htm

27. Peck R, Devore J. Roxy Peck's collection of classroom voting questions for statistics. 2008: Retrieved from http://mathquest.carroll.edu/resources.html

28. Utts J. Statistics 110/201 Practice Regression Final Key. 2009 Retrieved from http://www.ics.uci.edu/ jutts/110-201-09/PracticeRegressionFinalKey.pdf

29. Lewis A. Lecture 10: regression, correlation and acceptance sampling. 2012: Retrived from http://cosmologist.info/teaching/STAT/Statistics10.pdf

30. Department of Mathematics and Statistics, York University. (2005). Visualizing relations in multiple regression. Retrieved from http://www.math.yorku.ca/SCS/spida/lm/visreg.html

31. European Environment Agency. (2011, September 26). Illustration of the statistical analysis using multiple linear regression. Retrieved from http://www.eea.europa.eu/data-and-maps/figures/illustration-of-the-statistical-analysis

32. ASPH. Annual Data Report. (2010). Washington, DC: Association of Schools of Public Health, 2010. http://www.asph.org/UserFiles/DataReport2010.pdf