Volume 15 Number 3 September 2014

The Effect of Class Size on Reliability Estimates of College-Students Course Grades

Dr. Hassan G. AL-Omari
Department of Measurement and Evaluation
Faculty of Education -Jadara University
Hass_alomari@yahoo.com

Dr. Mutasem M. Akour
Faculty of Educational Sciences
The Hashemite University
Mutasem@hu.edu.jo

Volume 15 Number 3 September 2014

The Effect of Class Size on Reliability Estimates of College-Students Course Grades

Dr. Hassan G. AL-Omari
Department of Measurement and Evaluation
Faculty of Education -Jadara University

Dr. Mutasem M. Akour
Faculty of Educational Sciences
The Hashemite University

Abstract

This study aimed at investigating the effect of class size on reliability estimates of college-students course grades. Course grades were considered as composite scores with congeneric parts: first exam score, second exam score, final exam score, and attendance score. The reliability of these scores was estimated using Raju formula with three or more known lengths. To conduct this study, 63 classes were sampled from among all classes at Jadara University in Jordan in the second semester 20011/2012. These classes represented small, medium and big sizes with equal number of classes for each size. The results of this study showed that, in general, reliability estimates for all classes were low. The mean reliability estimate for all classes was 0.55 with 25% of classes being classified as having low reliability, and 65% of classes as having unacceptable reliability. The mean reliability estimate for small classes was 0.68, whereas it decreased to 0.41 for big classes. Finally, the relationship between class size and reliability estimates was shown to be significant, with small classes having higher estimates of reliability.

Key words: class size, Reliability, Raju coefficient, composite scores, course grades.

Received on: 28/2/2013 Accepted on: 29/9/2013

أثر حجم الشعبة الدراسية على تقديرات الثبات للعلامات الجامعية

د. حسان غازي العمري
قسم القياس والتقويم
كلية العلوم التربوية – جامعة جدارا

د. معتصم محمد عكور
كلية العلوم التربوية
الجامعة الهاشمية

الملخص

تهدف هذه الدراسة إلى الكشف عن أثر حجم الشعبة الدراسيةعلى تقديرات الثبات للعلامات الجامعية،حيث اعتبرت العلامة الجامعية التي يحصل عليها الطالب في نهاية المساق علامة مركبة من أجزاء متشاكلة (Congeneric) هي: علامةالامتحان الأول، الامتحان الثاني،الامتحان النهائي،وعلامة المشاركة، تم تقدير ثباتها باستخدام معادلة راجو لثلاثة أجزاء فأكثر معروفة الأوزان. لتحقيق أغراض الدراسة تم اختيار ٦٣ شعبة دراسية عشوائيا من الشعب الدراسية المطروحة في جامعة جدرا للفصل الدراسي الثاني ٢٠١١/٢٠١٢ بواقع ٢١ شعبة لكل من الشعب الصغيرة والمتوسطة والكبيرة .

أشارت نتائج الدراسة بشكل عام إلى انخفاض مستوى الثبات للعلامات الجامعية حيث وصل المتوسط العام إلى (٠,٥٥)، كما وبينت أن ما نسبته ٢٥٪ من الشعب الدراسية كانت منخفضة الثبات و٦٥٪ منها غير مقبولة الثبات. أما على مستوى حجم الشعبة الدراسية فتشير بيانات الدراسة إلى أن متوسط الثبات في الشعب الصغيرة (٠,٦٨)، فيما انخفض إلى (٠,٤١) في الشعب الكبيرة، كذلك أشارت النتائج بدلالة إحصائية أن الشعب الصغيرة تتمتع بثبات أعلى نسبيا.

الكلمات المفتاحية: حجم الشعبة، الثبات، معامل راجو، العلامة المركبة، العلامة الجامعية.

Volume 15 Number 3 September 2014

The Effect of Class Size on Reliability Estimates of College-Students Course Grades

Dr. Hassan G. AL-Omari
Department of Measurement and Evaluation
Faculty of Education -Jadara University

Dr. Mutasem M. Akour
Faculty of Educational Sciences
The Hashemite University

Introduction:

Class size is one of the important factors that affect students' scores, and evaluation tools implemented by instructors at colleges to assess their students. Many studies were conducted to address the effect of class size on students' achievement, class attendance, and the recommended methods and strategies for teaching. Reduction in class size resulted in better students' achievement (Achilles, 2003; Finn, 2002; Finn, Gerber, Achilles, & Boyed, 2001; Graue, Oen, Hatch, Rao, & Fadali, 2005; Smith, Molnar, & Zahorik, 2003). Moreover, reducing class size and using appropriate assessment tools affected students' outcomes positively and improved quality of education (Gibbs & Lucas, 1996; Graue & Ruscher, 2007). Small classes increase interaction between students and their instructors which enables instructors to have a better understanding of their students' strengths and weaknesses (Biddle & Berliner, 2002), and motivates students to attend classes and increase their in-class participation which results in improving their achievement (Bracey, 1995). Instructors in big classes are often engaged in the struggle to maintain discipline in class. This often leads to a reduction in teacher-learner contact for supervision and identification of learning difficulties in the learner (Molnar, et al. 1999).

National Council of Instructors of English (NCIE, 1997) pointed out that effective methods of teaching and assessment tools for small classes might be inefficient for larger classes. Most classes for freshmen are big due to the rising of education cost (Chapman & Ludlow, 2010). This leads to students being staked in small number of sections and classes, and thus limits the ability of instructors in having good communication with their students and in designing and using appropriate assessment tools. In addition, the

absence rate in big classes increases, obviously, after the first exam which threatens the consistency of students' performance and, consequently, the consistency of their grades (Honathan & Spence, 2010). This inconsistency leads to invalid interpretations of grades, and thus invalid decisions made upon these grades (Frisbie, 1988). Such inconsistencies and consistencies in students' performance are usually quantified through the estimation of a reliability coefficient of their grades (Feldt & Brennan, 1989).

Estimates of Reliability:

Reliability is one of the important psychometric properties of a test; it refers to the consistency of measurements (Traub, 1994). In other words, reliability is an indication of the accuracy of measurements which is better conceived in statistical terms through a statistical framework (Haertel, 2006). One of these notable frameworks is classical test theory, in which the observed score of an examinee on a given test form is the sum of a true score component and an error component (Haertel, 2006). In the estimation of internal consistency reliability, the total test (or the total score) must be decomposed into k separately parts. If X is the observed score for the total test, and X_1 … Xk are the part-test scores, then $X = X_1 + X_2 + \ldots X_k$ (Feldt & Brenan, 1989).

Three different models in part tests are distinguished in order to better understand various approaches to the estimation of reliability. These are called classically parallel, tau-equivalent parallel, and congenrically parallel (Feldt & Brennan, 1989). These models differ in the distributions of observed scores, true scores, and error scores on the parts of the test. Classically parallel parts assume that true-score variances and error score variances for the parts are equal. One of the well-known formulas of reliability that assumes classical parallel parts is the Spearman-Brown formula (Haertel, 2006). If the parts of a test can be assumed to meet the weaker assumptions of essentially tau-equivalence, part tests may exhibit differences in mean and differences in error score variances. Coefficient alpha and Kuder-Richardson formula 20 are based on the assumption that part tests are essentially tau-equivalent.

It should be noted that both the classical parallel model and the

essentially tau-equivalent model require the part tests to be equal in their functional length in order for true score variances to be assumed equal (Feldt & Brennan, 1989). For tests containing multiple item formats, or tests employing a single item type but but the weight of some items more heavily than others because of their importance, the separately scored parts may well vary in their functional length. Thus, part tests in these instances are unlikely to fit the essentially tau-equivalent model or the classically parallel model; It can only be modeled through the adoption of the congeneric model (Qualls, 1995).

For the congeneric model, true-score variances and error-variances may differ for each part. There is no unique solution for estimating reliability in this case. When the relative lengths of the subparts assumed to be known, Raju (1977) proposed the following formula,

$$R\rho_{XX'} = \left(\frac{1}{1-\sum\lambda_i^2}\right)\left(1-\frac{\sum\sigma^2{x_i}}{\sigma_x^2}\right) \quad (1)$$

Where λ_i is the relative length of part test i or the proportion of total test length for part test i, $\sum\lambda_i = 1$, and $\sigma^2{x_i}$ is the variance of each part test, and σ_x^2 is the composite score variance.

Composite Scores:

A composite score is any linear combination of two or more component scores, with fixed weights. The weights might be positive or negative and might be greater than, equal to, or less than 1.0, depending on the nature of the composite score (Feldt & Brennan, 1989). If XP1, XP2, …., XPk represent k component scores with weights w1, w2, …wk, a composite score for a person , Zp, may be represented as

$$Zp=w_0 + w_1Xp_1 + w_2Xp_2 + \ldots.. + w_kXp_k, \quad (2)$$

where w0 is an additive constant that may appear.

Course grades are examples of composite scores in the sense that they are an algebraic sum of two or more weighted scores: first exam scores, second exam scores, final exams scores, and scores assigned for assignments and attendance in some cases (Feldt, 2004). These parts can be considered congeneric because there is no guarantee that true score variances and

error score variances are equal since they will have different weights (Feldt & Brennan, 1989). In addition, the components of course grades result from using different tests administered under different conditions. Moreover, all of the previously discussed reliability estimates are considered as examples of composite scores subject to the constraint that all score components are at least congeneric (Heartel, 2006). As a result, Raju formula in equation (1) can be used in estimating reliability for composite course grades.

Some studies investigated the reliability of course grades and the effect of college type and level of study on reliability estimates. Sawalmeh (1995) found that 56% of courses at Yarmouk University had reliability estimates greater than 0.70, and course grades for colleges of economics and art tended to be more reliable than course grades for colleges of science and educational sciences. Alshayeb (2007) estimated the reliability of grades of 64 courses at Al-albayt University using Raju coefficient. It was found that, in general, reliability estimates were low; only 31.25% of courses had acceptable reliability estimates.

In addition, some studies (Bligh, 1988; Noble, 1991) pointed out that college grades have low reliability. This results from fact that instructors differ on how they assign grades. Some instructors are permissive, and some are stringent. These differences are due mainly to the difference in instructors' points of view and to their educational philosophy.

Class size reduction is a debated issue in education (Fisher et. al., 2001). Many studies investigated the relationship between class size and other variables such as students' achievement, students' attendance, and teaching methods used by instructors. However, this study is not mainly concerned with this debate in the sense that it is not a study about the relationship among these factors. The main concern of this study is the direct effect of class size on the consistency of course grades. And since reliability of grades reflects the degree of accuracy in measuring students' achievement, all factors that affect accuracy of measuring achievement are expected to affect grades reliability. One of these factors is class size. Therefore, the main focus of this study will be on estimating reliability for grades on different college courses and trying to relate the discrepancies in these estimates to the varying conditions of class size.

Statement of the problem:

Reliability is an important property of grades and it is an indication of the accuracy of measurements. Since reliability of grades reflects the degree of accuracy in measuring students' achievement, all factors that affect the accuracy of measuring achievement are expected to affect grades reliability. One of these factors is the inconsistencies on the part of those who evaluate examinee performance, i.e. the instructors (Feldt & Brennan, 1989).

Due to the rapid expansion in higher education across the world, student numbers have grown considerably in many courses, especially at the undergraduate level. The existence of large class sizes limits the ability of instructors in having good communication with their students, and in designing and using appropriate assessment tools which may threaten the consistency of students' grades. This inconsistency leads to invalid interpretations of grades, and thus invalid decisions made upon these grades (Frisbie, 1988).

Therefore, the main focus of this study will be on estimating reliability for grades on different college courses and trying to relate the discrepancies in these estimates to the varying conditions of class size. More specifically, this study aims at answering the following two questions:

1) What are the reliability estimates of college-students course grades?

2) What is the effect of class size on the reliability estimates of college-students course grades?

Significance of the study:

The significance of this study stems from the fact that based on the authors best knowledge, it is the first study to deal with the effect of class size on reliability estimates of composite scores. It is hoped that this study provides university administrators with information about the effect of class size on the reliability of students' grades, and thus on the nature of decisions that instructors take regarding students' successes and failures based on their grades. This is specially at this time when all universities are trying to adhere to the high standards of accreditation by reducing the number of students in classes to have small student-to-faculty ratios, which is one of the important facets in quality assurance.

Moreover, the importance of this study is built upon the importance of college grades for students. These scores form the basis for almost all decisions that are to be made about students. For example, accepting students in any program, or in any scholarship, or in any higher academic degree depends heavily on their grade point average or, in other words, on their grades in all courses.

Study Method:

Data:

The data for this study came from course grades for 63 classes, representing 3-credit-hour courses, chosen at random from among undergraduate classes at Jadara University in Jordan in the second semester 2011/2012. All classes were selected from five colleges (Science, Law, Arts, Economics, and Education) such that 12 classes were selected at random within each college, with the exception of the College of Economics where 15 classes were selected . Each class with less than 20 students was considered as small class, with more than 20 and less than 40 students was considered as medium class, and with more than 40 students was considered as big class (Bracey, 1995; Haris, 2007). The sample of this study was distributed evenly into three sizes: 21 small classes, 21 medium classes, and 21 big classes.

Analysis:

Raju coefficient, equation 1, was used in estimating reliability of course grades for all classes together and for different class sizes. At the undergraduate level, relative lengths are fixed: 20% for the first exam, 20% for the second exam, 10% for students' participation and attendance, and 50% for the final exam. Thus, $\lambda_1 = 0.20$, $\lambda_2 = 0.20$, $\lambda_3 = 0.10$,and $\lambda_4 = 0.50$. Then, for all classes, $\frac{1}{1-\sum \lambda_i^2} = 1.515$, The variance of each part test and the variance of the composite score were computed. In addition, descriptive statistics of Raju coefficients that were reported for all classes together and for different class sizes were also computed using SPSS 17.

Reliability estimates were classified into three categories: high, medium,

and low. No absolute standards are available to say whether a reliability estimate is high enough (Frisbie, 1988). If scores are to be used for individual assessment, reliability estimates more than 0.70 are considered to be high (Feldt & Brennan, 1989). However, reliabilities around 0.50 are considered to be acceptable for instructor made tests if scores will be combined with other information (such as: quiz scores, observations, etc.) to assign a grade for the course (Frisbie, 1988). Frary (2011) suggested a four-level classification of reliability coefficients. Reliabilities more than 0.90 are considered to be high, 0.80-0.89 are labeled as good, 0.60-0.79 as low to moderate, and 0.40-0.59 as doubtful. In the present study, reliability estimates were classified into three categories to be consistent with the classification of class size and to be consistent with the classifications of reliability estimates presented by Frisbie (1988) and Feldt and Brennan (1989). Based on these suggestions, reliability estimates equal to or higher than 0.70 were labeled as high, between 0.40 and less than 0.70 were labeled as medium, and less than 0.40 were labeled as low. In addition, reliability estimates less than 0.70 were labeled as unacceptable.

Results:

The purpose of this study was two-fold. To examine the reliability estimates of students' grades at the college level, and to investigate the relationship between class size and reliability estimates of students' grades. To achieve this purpose, Raju coefficient was used to estimate reliability of students' grades in 63 classes taken at random from among the classes at one of the universities in Jordan, Jadara University. For the resulted reliability estimates, descriptive statistics were computed for all classes and for each class size and displayed in Tables 1 and 2.

Table 1 shows descriptive statistics for reliability estimates of all classes taken together. The distribution of these estimates was negatively skewed with mean 0.55 and variance 0.08. The maximum value of reliability estimates was 0.99 and the minimum value was -0.13. Twenty two out of 63 classes (35%) were classified as having high reliability estimates, twenty five classes (40%) as having medium reliability estimates, and sixteen classes (25%) as having low reliability estimates. As a total, forty one classes (65%)

were classified as having unacceptable reliability estimate.

Table 1
Descriptive statistics of reliability estimates of all classes (63 classes)

Statistic	Value
Mean	0.55
Variance	0.08
Skewness	-0.69
Maximum	0.99
Minimum	-0.13
Range	1.12
Number of classes with high reliability estimates	22
Number of classes with medium reliability estimates	25
Number of classes with low reliability estimates	16
Number of classes with unacceptable estimates of reliability	41

At the class size level, Table 2 shows that the average reliability estimates was 0.68 for small classes, 0.51 for medium classes, and 0.41 for big classes. Reliability estimates ranged from 0.22 to 0.94 in small classes, from - 0.12 to 0.90 in medium classes, and from - 0.13 to 0.73 in big classes. The variance of reliability estimates for each class size indicated that these estimates were more homogenous for small classes as compared to medium or big classes. This is also evident when comparing the number of classes with high reliability estimates for different sizes.

Moreover, it can be seen from Table 2 that the number of classes with unacceptable reliabilities increased as class size increased. Nineteen big classes (91%) had unacceptable reliabilities as compared to 13 (62%) medium classes and 9 (43%) small classes.

Table 2
Descriptive statistics of reliability estimates according to class size

Statistic	Class size		
	Small	Medium	Big
Mean	0.69	0.51	0.45
Variance	0.05	0.11	0.06
Skewness	-0.89	-0.57	-0.88
Maximum	0.99	0.90	0.73
Minimum	0.22	-0.12	-0.13
Range	0.77	1.02	0.86

Table 2 Countied

Statistic	Class size		
	Small	Medium	Big
Number of classes with unacceptable reliability estimates	9	13	19

Figure 1 shows that more small classes had high reliabilities as compared to medium and big classes. In addition, Figure 1 and Table 2 show that 12 small classes had high reliability estimates as compared to 8 medium classes and 2 big classes. The relationship is reversed when comparing the number of classes that had low reliability estimates.

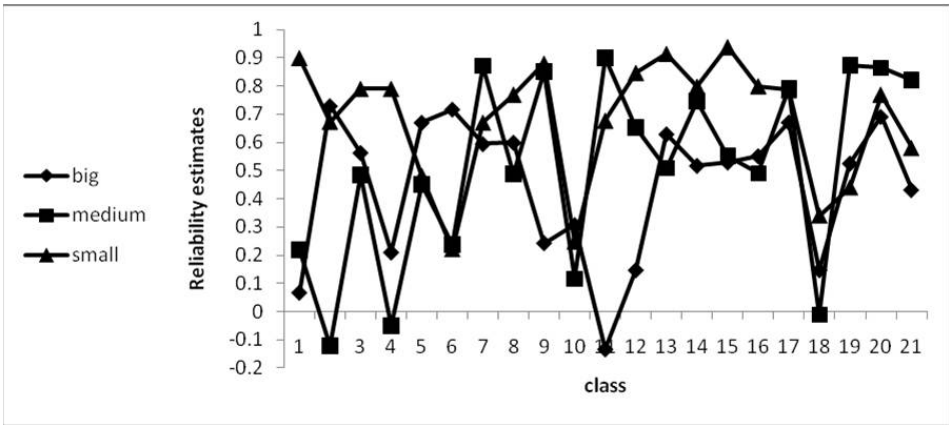

Figure 1
Reliability estimates for different classes at each size level

In order to examine the relationship between class size (small, medium, and big) and estimates of courses-grades reliability (low, medium, and high), the number of classes for each class size was computed for the three levels of reliability estimates. Table 3 shows number of classes in each cell that resulted from the crossing of the two factors: class size and estimates of reliability.

Table 3
Number of classes for each condition of class size and reliability estimates

Reliability estimates	Class size		
	Small	Medium	Big
High	12	8	2
Medium	6	7	12
low	3	6	7

Table 3 shows that the number of classes with high estimates of reliability increased as class size decreased. Out of 21 small classes, 12 (57%) classes exhibited high reliability estimates. However, this percentage declined to 38% (8 classes) for medium classes and to 10% (2 classes) for big classes. On the other hand, the number of classes with medium and low estimates of reliability increased as class size increased. Out of 21 classes, 3(14%) small classes had low reliabilities as compared to 7 (33%) big classes.

In order to test the significance of the relationship between class size and reliability estimates, Chi-square test of independence or relatedness was conducted on the cells in Table 3. The observed test value was ($2\square$ = 11.01, df = 4, p-value = 0.026) which was significant at the 0.05 level. Since chi-square test reveals the significance of the relationship and does not show the strength and magnitude of the relationship, the contingency coefficient was computed for the data. The strength of the relationship using the contingency coefficient was 0.38 with p-value= 0.02 which was significant at the 0.05 level.

Discussion and conclusions:

The results of this study showed that, generally speaking, reliability estimates of university grades were low. This agrees with previous studies (Bligh, 1988; Noble 1991); the average reliability estimates was 0.55 for all classes with a variance of 0.08, which indicates the clustering of reliability estimates around the mean value. This may be due to the fact that most instructors do not have enough knowledge and training on applying appropriate tools of assessment and on building and conducting achievement tests with good psychometric properties. Therefore, it is

necessary to provide instructors with sufficient training on those issues in order to be more consistent in assigning grades to students. Increasing the reliability of course grades will build more trust in the decisions that are to be made upon these grades.

Reliability estimates ranged from, surprisingly, – 0.13 in big classes to 0.94 in small classes. The presence of such extreme estimates was reflected in the value of the range, a value of 1.07. Negative estimates occur when there are negative covariances between part tests that constitute the composite score. One possible reason for this discrepancy is that classes in this study were sampled from different colleges at the university; this agrees with the findings of Alsawalmeh (1995). Another possible reason that explains this discrepancy is class size; the results of this study revealed that negative reliability estimates existed in medium and big classes, with more variability in reliability estimates as compared to small classes.

Furthermore, the results of this study showed that 35% of all classes had acceptable and high reliability estimates, which represents only one third of the entire group of classes. On the other hand, 25% of classes had low reliabilities. This result is worthy of noting because what we are saying here is that about a quarter of all courses were not measuring students performance accurately and all future decisions and interpretations based on these measurements are questionable.

These findings are supported by results at the class size level. Small classes had relatively high estimates of reliability as compared to medium and big classes. Average reliability estimates was 0.68 for small classes, 0.51 for medium classes, and 0.41 for big classes. In addition, the percentage of courses with unacceptable estimates of reliability increased from 9 for small classes to 19 for big classes. This relationship between class size and reliability estimates showed to be significant, which indicates that measurements in big classes are more error prone. Thus, it would be more difficult for instructors to defend their decisions about students' successes and failures. These findings agree in part with the findings reported by other researchers (Achilles, 2003; Finn, 2002; Finn et al., 2001; Graue et al., 2005; Gibbs & Lucas, 1996; Smith et al., 2003) in that students' achievement and communication with instructors are better in small classes.

It is highly recommended that university administrators take class size into consideration as an important factor that affects the quality of education. Students have the right to have good teaching, the opportunity to interact with the instructor and with other students, and grades that truly reflect their competences.

One of the limitations of this study is that classes were sampled from one university in Jordan, Jadara University. The results may not be generalizable to other universities with different assessment systems and with instructors having different teaching abilities as compared to those at Jadara University. Therefore, it is recommended to conduct more research using representative sample that includes more and differentiable universities. It is also recommended to sample classes with different sizes for the same instructor and, then, investigate the relationship between class size and reliability when instructors were held as constant. This might have a better insight into the effect of class size on reliability.

Reliability is a necessary ingredient of validity, but it is not sufficient to insure validity (Frisbie, 1988, pp100-101). Therefore, it is recommended to investigate the effect of class size and other factors related to students learning (teaching methods, student motivation, etc.) on the validity of course grades. Finally, decision makers at universities should pay more attention to class size in the sense that larger class sizes pose significant teaching challenges, not least in the assessment of student learning. The increase in class size poses certain constraints on designing manageable and yet effective forms of assessment that possess an acceptable indices of reliability and validity. It is recommended to train instructors on all issues that relate to the accuracy of measuring students' outcomes. This will help in decreasing errors in assigning grades to students, and thus increasing the reliability of these grades, which in turn will enhance the validity of the meaning of these scores.

References:

Achilles, C. M. (2003). *How class size makes a difference: What the research says about the impact of size reduction*. A Paper presented at the SERVE Research and policy symposium on class-size reduction and beyond, February 24, USA: Raleigh.

Al Shayeb, A. (2007). Estimating reliability for courses grades at Al Albayt University. *Damascus University Journal, 23*(2), 271-255.

Biddle, B. J., & Berliner, D. C. (2002). Small class size and its effects. *Educational Leadership, 59*(5), 12-23.

Bligh, D. (1988). *Higher education*. London: Cassel Educational limited.

Bracey, W. (1995). Research oozes into practice: the case of class size. *Phi Delta Kappa, 77,* 89–90.

Chapman, L., & Ludlow, L. (2010). Can downsizing college class sizes augment student outcomes? An investigation of the effects of class size on student learning. *The Journal of General Education, 59*(2), 105-123.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in Counseling and Development, 37*(3), 184-190.

Feldt, L. S., & Brennan, R. (1989). Reliability, in R.Linn (Ed), *Educational Measurement* (3rd ed., pp. 105-146). New York: Macmillan Publishing Company.

Finn, J.D. (2002). Small classes in American schools: Research, practice and politics. P*hi Delta Kappa, 83*(7), 551-560.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Instructors College Record, 103*(2), 145-183.

Fisher, D., Lapp, D., Flood, J., Frey, N., & Moore, K. (2001). Why class size matters: An investigation of teacher, administrator, and parent perceptions. *National Reading Conference Yearbook, 50,* 189-199.

Frary, R. (2011). *Testing Memo 8: Reliability of test scores*. Retrieved February 12, 2012, from http://www.testscoring.vt.edu/memo08.html.

Frisbie, D. (1988). Reliability of scores from instructor-made tests. *Educational Measurement: Issues and Practice, 7*(1), 25-35.

Gibbs, G., & Lucas, L. (1997). The effects of class size and form of assessment on nursing students' performance: approaches to study and course perceptions. *Nurse educational today journal, 17*(4), 311-318.

Graue, E., Oen, D., & Rauscher, E. (2007). *Understanding how class size reduction & assessment shape education experiences*. Wisconsin Center for Education Research: University of Wisconsin Madison. Retrieved March 8, 2012 from http://varc.wceruw.org/sage/Class%20size%20reduction%20and%20assessment%20final.pdf.

Graue, E., Oen, D., Hatch, K. Rao, K., & Fadali, E. (2005). *Perspectives on class size reduction*. A paper presented at the annual meeting of the American Educational Research Association, April 12, Montreal, Canada.

Haertel, E. H. (2006). *Reliability. in R. Brennan (Ed), educational measurement* (4th ed, pp. 65-110). New York: American Council on Education and Macmillan.

Harris, K-L., Krause, K., Gleeson, D., Peat, M., Taylor, C. & Garnett, R. (2007). *Enhancing Assessment in the Biological Sciences: Ideas and resources for university educators*. Retrieved March 1, 2012, from www.bioassess.edu.au.

Cole, J., & Spence, S. (2010). *First year fluids – encouraging student engagement when the class size is large*. A paper presented at the 3rd International symposium for engineering education, July, Cork, Republic of Ireland.

Molnar, A., Smith, P., Zahovik, J., Palmer, A., Halback, A. & Ehrie, K. (1999). Evaluating the SAGE Program: A Pilot Program in Targeted Pupil-Teacher Reduction in Wisconsin. *Education Evaluation and Policy Analysis, 21*(2), 165–77.

National Association of Secondary School Principals. (1996). Breaking Ranks: Changing an American Institution. *NASSP Bulletin, 80*(578), 55- 64.

National Council of Instructors of English. (1997). *More than a Number: Why Class Size Matters*. Council-Grams: News and Information for Leaders of the Council (Vol. LX, No. 2). Urbana, IL: National Council of Instructors of English.

Noble, J. (1991). *Predicting college course grades using ACT assessment scores and high school course work information*. (ACT research report series, 91-3). Iowa City, IA: American College Testing.

Qualls, A. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8*(2), 111-120.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549-565.

Sawalmeh, Y. (1995). Estimating reliability of grades for a sample of courses at Yarmouk University in the second semester 1992/1993. *Journal of the E. R. C., 4*(7), 71-89.

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*(4), 419-433.

Smith, P., Molnar, A., & Zahorik, J. (2003). Class-size reduction: A fresh look at the data. *Educational Leadership, 61,* 72-74.

Traub, R. E. (1994). *Reliability for the social sciences. Theory and Applications*. London: Sage Publications, Inc.