The Reliability of Complex Item Types for Assessing Cognitive Ability

Charles Secolsky'
Center for Instructional Research and Curriculum Evaluation
csecolsky@gmail.com

The term 'reliability' is most often used as a statistical estimate defining error associated with the administration and scoring of parallel forms of a test, error associated with the re-administration of and scoring of the same test form over occasions, the internal consistency of scores on a form from one administration, or even the error associated with the scores from different judges. However, little has been done methodologically to investigate the reliability of two different tasks which are part of the same assessment form. The present paper is concerned with just that –estimating the reliability of two different tasks associated with the same form. This paper presents methods of reliability for three different complex innovative item types where the actual item is one component and the other components are: (1) topics associated with the item, (2) solution strategies (intrinsic difficulties) associated with answering the question, and (3) conceptual distances between topics. Reliability was achieved for types (2) and (3) where the associated tasks were uncorrelated with responses to the actual items. Generalizability theory was used for type 1 but correct answers to the items were associated with choosing the topic selected by the instructor.

The Reliability of Complex Item Types for Assessing Cognitive Ability

Recent concerns regarding the assessment of examinee response processes via collecting additional item level data for validating arguments of test score interpretations have resulted in a renewed interest in using complex item types. These items are intended to ensure the identification of faulty test questions prior to identifying them as faulty at the item and test analysis phases as construct-irrelevant variance. Such item types consist of ancillary procedures which ask examinees to indicate the topic/concept associated with each item, providing additional multiple-choice options representing examinee misconceptions that may be unrelated to the construct measured; and, differences between examinee and instructor conceptual distances of the topics/concepts via multidimensional scaling (MDS). These items are more complex than the generic breed of objective or free response items since their administration demands scores for each item that are beyond correct and incorrect responses. They also demand that differences between examinee and instructor judgments of topics, examinee selected misconceptions in the form of *intrinsic difficulties*, and concept mapping using MDS be part of the analysis. While their usefulness has been demonstrated on a limited basis, no attempts have been made to estimate reliability of such ancillary complex item types, which is the purpose of this study.

The author proposes that reliability be estimated using a generalizability framework for examinee and instructor judgments of topics/concepts where the total variation of scores and total test judgments are analyzed. A G-study consisting of examinee item scores, examinee-selection of instructor indicated topics and examinee selected topics irrespective of topics selected by instructors and interactions terms is proposed. For internal consistency of intrinsic difficulties, the item scores (correct or incorrect) are correlated with selected misconceptions for determining whether the four misconceptions per item are uncorrelated with actual scored responses. For conceptual distances, reliability can be estimated using reliability of item and distance measure composites.

Generalizability of Examinee and Instructor Topic/Attribute Selection

The method used to assess the generalizability of the complex item type that is concerned with the collection of additional data in the topic selection for instructors and examinees requires that instructors as the experts choose the most appropriate set of topics from the list of topics that have previously been used to represent the topics corresponding to the items. The approach used is a G study whereby item scores of examinees, topic scores of instructor selected topics, examinee selected topics, and the interaction terms are used to generate variance components in generalizability or G studies. The proportions of variation attributable to each of these effects define the robustness of this type of innovative item type. When the interaction terms between instructor selected topics and examinee related topics accounts for a relatively small component of variation, then it can be stated that this form of reliability (generalizability) for this type of innovative item type is the gold standard of high in reliability. Alternatively, if each variance component is converted to a proportion which adds to 100%, then the coefficient of reliability is defined for this interaction term. As an example, consider the 20 item test of a basic skills examination with 13 topics/attributes. While the presence of the topics and the process of their selection are intended for aiding in the validation of the items, a sense of their reliability provides

for their stability with respect to whether the examinees understood the topic selection task. The topics used for this study were:

a) Reducing Fractions
b) Dividing Fractions
c) Multiplying Fractions
d) Division of Common Factors
e) Multiplying by the reciprocal
f) Adding Fractions
g) Finding the Least Common Denominator
h) Subtracting Fractions
i) Converting mixed numbers to improper fractions.
j) Finding the correct place value.
k) Converting fractions to decimals
l) Multiplying a decimal fraction by 100.
m) Changing a fraction to a percent

The equations for the G-study are as follows from Webb, Shavelson and Steedle (2012).

$X_{stj} =$ (1)

μ	grand mean
$+ \mu_s - \mu$	total item score effect
$+ \mu_t - \mu$	instructor-selected topic effect
$+ \mu_j - \mu$	examinee-selected topic effect
$+ \mu_{st} - \mu_s - \mu_t + \mu$	score × instructor-topic effect
$+ \mu_{sj} - \mu_s - \mu_j + \mu$	score × examinee-topic effect
$+ \mu_{tj} - \mu_t - \mu_j + \mu$	instructor × examinee topic effect
$+ X_{stj} - \mu_{st} - \mu_{sj} - \mu_{tj} + \mu_s + \mu_t + \mu_j - \mu$	residual/error

$$\sigma^2_{X_{stj}} = \sigma^2_s + \sigma^2_t + \sigma^2_j + \sigma^2_{st} + \sigma^2_{sj} + \sigma^2_{tj} + \sigma^2_{stj,e}.$$

An Empirical Example of the Generalizability of the Innovative Item Type with Instructor-Examinee Topic Selection.

Table 1 below provides the variance component estimates for the total score following the model above, the topic scores, and each of the selected topic (only the first indicated topic was used in the calculation), and the percent of total variability. Table 2 provides the estimates after summing over all the judgment scores (J1–J20). Note that the judgment scores were not included in any interaction terms because of the already complex G- study used.

Table 1. Variance Component Estimates for @@@

Source	Estimate	Percent Total Variability
Total Score	79.69	5.29
Topic Score	0	0
J1	0	0
J2	0	0
J3	0	0
J4	59.09	3.93
J5	0	0
J6	97.62	6.49
J7	120.83	8.28
J8	105.19	7.0
J9	0	0
J10	54.31	3.61
J11	190.31	12.64
J12	27.98	1.86
J13	99.81	6.63
J14	8.07	0.54
J15	0	0
J16	144.10	9.57
J17	1.16	0.07
J18	0	0
J19	22.75	1.51
J20	0	0
Total Score x Topic Score	299.11	19.87
Error	195.06	12.96

Table 2. Variance Component Estimates Summing Over Judgment Scores

Source	Estimate	Percent Total Variability
Total Score	79.69	5.29
Total Judgment Score	1505	61.3
Total Score x Topic Score	299.11	19.87
Error	195.06	12.96

As can be seen in Table 2, approximately 20% of the total variability is attributable to the Total Score x Topic Score interaction. The total judgment score variability is 61.3 %. It will be assumed here that the Total Score x Topic Score interaction is a measure of the independence of the topics as they relate to the total scores. In the present study, this borders on being problematic since indicating the instructor-selected topic is to some extent related to obtaining the correct answer to a question.

The Reliability of Intrinsic Difficulties

Intrinsic difficulty represents what examinees find difficult in items or text. Magaram, Phanor Secolsky, and Hasbrouck (2011) first translated student misconceptions in think-aloud protocols (see Ericsson & Simon, 1993). Intrinsic difficulty can be differentiated from proportion correct scores or p-values since it is dependent on what students express in taped interviews on how to solve fraction and decimal computation problems as was the case in the Secolsky, Kossar, Magaram and Fuentes (2011) study. In essence, students answered a series of questions. The responses were scored as correct or incorrect. In addition, students selected from four choices as many of the choices that they considered correct solution strategies. So for each student in this complex innovative item type, there are two scores. One is a vector of correct and incorrect responses for the 20 item test. Also, there is a set of selections: a, b, c, and/or d for the four solution strategies most of them representing misconceptions or incorrect solution strategies.

The reliability that would seem to have the greatest utility is actually not the consistency of the solution strategies but rather the correlation of the correct versus incorrect weighted responses by the frequency of the selection of solution strategies (each of the four ancillary options per item). If the correlation is very high it likely means that correct and incorrect responders are selecting the same misconceptions (solution strategies). Otherwise, if this correlation is low or the frequencies are virtually uncorrelated, then the misconceptions are not differentiating high and low abilities on each of the 20 individual items. Two sample items are shown below in Figure 1.

These items taken from the assessment form with 20 such complex item types are scored in the following manner. Students receive a 1 for a correct response and a 0 for an incorrect response to the actual item on the left side and choose from a list of four solution strategies those strategies that represent correct strategies to each examinee.

Item	Response Options
1) Divide and simplify $\frac{7}{4} \div 7$	a) You would have to multiply by ¼. So you get 7/4. 7/4 divided by 7/4 equals 1. b) You start out by changing 7/4 to 1 ¾ and then dividing by 7. c) You should start out by changing to a multiplication problem: 7/4 times 1/7. d) After you have 7/4 times 1/7 you cross multiply to get 49/4 or 12 ¼.
2) Add and simplify. $\frac{7}{9} + \frac{5}{6}$	a) I add the numerators and add the denominators to get 12/15. Then I simplify to get 4/5. b) First, I find the lowest common denominator by multiplying 9 by 6 =54. c) The lowest common denominator is 18. I then multiply 2 by 7 and 3 by 5 = 14+ 15= 29/18 = 1 11/18. d) The lowest common denominator is 36. 28/36 + 30/36 = 58/36 = 1 22/36 = 1 11/18.

Figure 1. Sample items with response options

An Empirical Example of the Reliability of Intrinsic Difficulties

A test containing 20 of these innovative item types was administered to 238 basic mathematics students at a New Jersey high school. The range of total scores on the actual items was 0–13. The range of total strategy scores also ranged from 0–13. Means, standard deviations for these two distributions as well the Pearson correlation coefficient between these two variables is presented in Table 3.

Table 3. Means, Standard Deviations and Pearson Correlations of Total Test Score and Total Solution Strategy Score

	Mean	Standard Deviation
Total Test Score	4.605	3.072
Total Strategy Score	5.815	3.089
Pearson r: r = 0.06430 (p = 0.323)		

As shown in Table 3, the value of correlation between the total solution strategy scores and the total test scores is nearly uncorrelated showing that in this testing, the ability to solve an item correctly is not associated with the ability to determine the correctness of the different

solutions strategies for items. Reliability as defined here seems to be at an acceptable level because it shows the independence of item scores and selected solution strategies.

The Reliability of Categorical Conceptual Distances as an Innovative Item Type

MDS has been used by Secolsky, Magaram, Arvanites and Levy (2013). For a 14 item test on statistics with 10 topics, there were 45 non-identical unique pairs of concepts in which students indicated on a scale of 1 to 3 the conceptual distance between topics. The rating of "1" was meant for a pair of topics that were similar to each other; the rating of "2" was meant for topics that were somewhat similar to each other; and the rating "3" was intended for a pair of topics that were perceived to be different from each other. From these data, an MDS map was produced from a distance matrix. Estimating the reliability of this complex item type was achieved by estimating reliability simplified from a more complex procedure devised for ordinal data as originally conceived by Zumbo, Gadermann and Zeisser (2007). Essentially, it is based on Cronbach's (1951) alpha but is the correlation between the points awarded for each of 14 items and the distance ratings by examinees between pairs of topics.

An Empirical Example of the Reliability Based on Total Item Score and Measures of Examinee Distances

As part of the innovative test design employed by Secolsky et al. (2013), 14 items worth varying numbers of points and 45 examinee judged distances between pairs of unique topics, Cronbach's (1951) alpha for the 14 items worth 100 points 0.856. When the 45 distance measures were submitted to the same SAS reliability program (PROC CORR with an alpha option), the obtained value of Cronbach's (1951) alpha coefficient was 0.844. When all 59 variables as items and tasks were computed together as a sort of composite measure, the resultant alpha was 0.849. These results imply that the addition of the 45 distance measures were likely independent of the earned point values for the test. To check on this finding, the sum of the points over items and the sum of the distances measures over pairs of topics were correlated. With n=50 students with non-mission data on the distance measure, r=0.021 (p=0.886). The result indicates that for this type of innovative item type, the item point values are independent of the distance measures. The near zero correlation supports the contention that their distance measures are providing unique information.

Discussion

The reliability of three different item types were explored as ancillary measures in addition to the actual item scores: (1) examinee identification of the instructor-selected topic, (2) intrinsic difficulties or solution strategies for each item consisting of four potential examinee misconceptions, (3) distance measures between concepts. The second and third ancillary procedures showed no relationship between examinee ability in the content area and the ability to answer the ancillary item type in a way that is systematically related to answering the actual question. For these two procedures the ability and judgment of distance provide different sorts of information. However, the ability to choose the instructor-selected topic was somewhat related to answering the actual item correctly.

The type of reliability selected for each of the three procedures was different. For the first ancillary procedures, a generalizability G-study formulation was devised. For the second ancillary procedure, a Pearson correlation was used. For the third procedure, the distance measure, Cronbach's alpha and Pearson correlation was used. In each of these applications, the goal was not to find out traditional reliability in the form of internal consistency, but rather whether the information contained in the ancillary measure provided unique information for use in finding faulty test questions prior to traditional item analyses even with pretesting.

One other concern about the reliability of ancillary procedures that need to be addressed is in the order of the administration itself. In what order should the actual test items and the ancillary tasks be presented to examinees. The order selected in these studies was strictly based on logic. Presenting the actual items first conceivably gives the examinees sufficient familiarity with the item they are later asked questions about (Secolsky, 1980).

References

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

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: verbal reports as data* Boston, MA.: MIT Press,

Magaram, E., Phanor, J., Secolsky, C. & Hasbrouck, P. (October, 2013). *Eliciting student judgments of intrinsic difficulty for studying student misconceptions in solving basic mathematics items* Paper presented at the Northeastern Educational Research Association.

Secolsky, C. (1980). *Assessing the interpretive component of criterion-referenced test item validity.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Secolsky, C. (1983) Using examinee judgments for detecting invalid items on teacher-made criterion-referenced tests. *Journal of Educational Measurement, 20,* 51-63.

Secolsky, C., Kossar, B., Magaram, E. & Fuentes, V. (October, 2011). *Estimating examinee intrinsic difficulty for providing greater specificity of feedback for instruction*. Paper presented at the annual meeting of the International Association of Educational Assessment, Manila, Philippines.

Secolsky, C., Magaram, E., Arvanites, P. & Levy, S. (April, 2013). *Improving validity by assessing students' thought processes and perceived conceptual distances on classroom tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Webb, N. M. Shavelson, R. J. & Steedle, J. T. (2012) *Generalizability theory in assessment contexts*. In C. Secolsky, & D. B. Denison (Eds.) *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132-149).

Zumbo, B., Gadermann, A. M., Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6(1),*21-29.