A Generalizability Approach to the Measurement of Score Reliability of the Teacher Assessment Literacy Questionnaire

Hussain Alkharusi

College of Education, Sultan Qaboos University

P.O.Box: 32 Al-Khod, P.C.: 123, Sultanate of Oman

Tel: 968-9622-2535 E-mail: hussein5@squ.edu.om

Received: March 11, 2011 Accepted: March 28, 2012 Published: May 1, 2012

10.5296/jse.v2i2.1495 URL: http://dx.doi.org/10.5296/jse.v2i2.1495

Abstract

Classroom assessment is one of the main responsibilities of the teachers. Sound classroom assessment practices require teachers to have adequate levels of knowledge and skills in the educational assessment. The *Teacher Assessment Literacy Questionnaire (TALQ)* was developed by Plake and Impara (1992) to measure teachers' knowledge and understanding of the basic principles of the educational assessment. This study applied generalizability theory to estimate the dependability of the TALQ's scores for pre-service teachers (N = 259) enrolled in an educational measurement course at Sultan Qaboos University in Oman. Results showed that the scores are highly generalizable across items with relatively small item variance components. Implications are discussed in relation to the reliability theory.

Keywords: Generalizability theory, Score reliability, Teacher Assessment Literacy Questionnaire, Educational assessment

1. Introduction

Classroom assessment refers to the process used in the classroom by the teacher to obtain information about students' performances on assessment tasks using a wide range of assessment methods to determine the extent to which students are achieving the target instructional outcomes (Gronlund, 2006). It is one of the main responsibilities of the teachers (Alkharusi, 2009). Sound classroom assessment practices require appropriate levels of assessment literacy, which is defined as "the ability to understand the different purposes and types of assessment in order to select the most appropriate type of assessment to meet a specific purpose" (Ainsworth & Viegut, 2006, p. 53). According to the "Standards for Teacher Competence in Educational Assessment of Students" set by the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) in 1990, assessment literacy entails teachers' ability to develop various assessment methods such as paper-pencil tests and performance measures; administer, score, and interpret assessment results; develop grading procedures; communicate assessment results; and use them in making educational decisions.

Recently, Brookhart (2011) argues that the "Standards for Teacher Competence in Educational Assessment of Students" do not consider current conceptions of formative assessment knowledge and skills and teacher's assessment literacy required to successfully work in the accountability and standards-based assessment context. As such, she proposed a set of educational assessment knowledge and skills for teachers in reference to formative assessment and standards-based assessment systems. The set states that teachers should:

1. Understand learning in the content area they teach.

2. Be able to set and apply the learning intentions that are congruent with both the content and depth of the standards and curriculum goals.

3. Have a repertoire of strategies for communicating to students about the expectation of the learning intentions.

4. Understand the purposes and uses of the various types of assessment and be able to use them.

5. Be skillful in analyzing classroom assessment methods.

6. Be skillful in providing effective meaningful feedback on student work.

7. Have the ability to develop scoring schemes to quantify student performance on classroom assessments conducive for making informed educational decisions.

8. Be skillful in administering external assessments and interpreting their results for decisions making.

9. Be able to apply their educational decisions made out from classroom assessments.

10. Be able to communicate assessment information to students to motivate them to learn.

11. Understand the legal and ethical issues in the classroom assessment practices.

Recognizing the need for teachers to possess an adequate knowledge and skills in classroom assessment, Plake and Impara (1992) developed an instrument titled the "*Teacher Assessment Literacy Questionnaire* (TALQ)" consisting of 35 items to measure teachers' assessment literacy. The TALQ was based on the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, & NEA, 1990). The instrument was administered to a sample of 555 in-service teachers around the United States. The results indicated that the teachers might not be well prepared to assess student learning as revealed by the average score of 23 out of 35 items correct (Plake, Impara, & Fager, 1993). Likewise, Campbell, Murphy, and Holt (2002) applied the TALQ to a sample of 220 undergraduate students who a completed a course in tests and measurement. The results revealed that the average score for the sample was 21 out of 35 items correct, suggesting the need for more attention to the assessment literacy of the prospective teachers. Recently, Alkharusi (2011a, 2011b) examined the psychometric properties of the TALQ for pre-service teachers in Oman. The results showed that the TALQ measures a unitary construct of the assessment literacy with an adequate level of internal consistency reliability, and that the items demonstrated acceptable levels of difficulty, discrimination, reliability, and validity.

Although the aforementioned studies demonstrated the utility of the TALQ in providing valid and reliable score interpretations of the assessment literacy, his approach was based on the classical theory (CT) of reliability which provides limited information by considering only one source of measurement error at a time (Alkharusi, 2012). Unlike the CT of reliability, the generalizability theory (GT) recognizes that multiple sources for error (i.e., error variance attributable to items or occasions) may occur simultaneously in the measurement process (Crocker & Algina, 1986; Shavelson & Webb, 1991). In the GT, the multiple sources of error are estimated using an analysis of variance model. These sources of error are called “facets”, which represent unwanted variation in the observations attributable to items or occasions. The GT estimates the variance components related to these sources to identify the optimal conditions needed to obtain reliable scores for high-stakes performance decisions, such as decisions regarding teachers’ assessment literacy level.

2. Purpose of the Study and Research Question

Given the growing interest on teachers' assessment literacy, research is needed to support the assertion that score interpretations from the available instruments such as the TALQ are dependable when making decisions regarding teachers' assessment literacy. Therefore, the purposes of the current study were (a) to explore the sources of variance for TALQ’s scores using generalizability theory and (b) to estimate the generalizability coefficient of the TALQ’s scores. The study was guided by the following research question:

To what extent are item scores on the Teacher Assessment Literacy Questionnaire dependable to draw inferences about teachers’ assessment literacy levels?

3. Methods

3.1 Sample

The sample for this study included 259 Omani pre-service teachers enrolled in a required undergraduate level educational measurement course in the College of Education at Sultan Qaboos University. There were 142 females and 117 males in the study. The participants ranged in age from 19 to 25 years with a mean of 21.08 and a standard deviation of 1.26. The following majors were represented in the sample: English language (*n* = 140), science and math education (*n* = 70), and educational technology (*n* = 49).

3.2 Setting

The undergraduate level educational measurement course is offered by the Department of Psychology in the College of Education at Sultan Qaboos University. The goal of the course is to have pre-service teachers develop knowledge, skills, and abilities related to classroom assessment that deemed essential to the teaching profession. The course is a three credit hour required course for all undergraduate education majors. A prerequisite for students enrolled in this course is to have completed and passed a course in educational objectives. Topics covered within the undergraduate level educational measurement course are basic concepts and principles in measurement and evaluation, teacher-made tests, standardized tests, test and item analysis, reliability and validity, performance assessment, grading, reporting, and communicating assessment results.

3.3 Instrumentation

In addition to the biographical information collected in terms of self-reported gender, age, major, and university identification number; the primary instrument in this study was the *Teacher Assessment Literacy Questionnaire* (TALQ) developed by Plake and Impara (1992) to measure teachers' knowledge and understanding of the basic principles of the classroom assessment practices. It consisted of 35 multiple-choice items with four options, one being the correct answer. The items were developed to align with the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, & NEA, 1990). The items were dichotomously scored (0 = *incorrect response*, 1 = *correct response*) with a high total score reflecting a high level of assessment literacy. Plake, Impara, and Fager (1993) reported a KR20 reliability coefficient of .54 for the in-service teachers' scores whereas Campbell, Murphy, and Holt (2002) reported a KR20 reliability coefficient of .74 for the pre-service teachers' scores. The TALQ has been validated for use with pre-service teachers in Oman by Alkharusi (2011a, 2011b). Internal consistency reliability for the TALQ's scores was found to be .84 as measured by KR20 reliability coefficient (Alkharusi, 2011a, 2011b).

3.4 Procedures

Instructors' permission to collect data from the students in this study was requested and obtained. At the end of a regularly scheduled class meeting, the author informed the students that a study was being conducted on the assessment literacy of the pre-service teachers. At this time, the author requested the participation of the students. Emphasis was placed on the fact that information to be gathered would not influence their course final grade in any way and that the study would hopefully lead to improved instruction in the course. Students who wished to participate in the study were provided with the TALQ, which also included the brief biographical information sheet. Total time for completing the instrument averaged approximately 90 minutes.

3.5 Statistical Analyses

The study employed a one-facet design of the generalizability analysis because the items facet is the only facet of potential measurement error for TALQ's scores. Following Alkharusi's (2012) illustration of the generalizability analysis, mean squares and variance components estimates were obtained using Type I method and a random effects model. The analysis is based on the assumption that the levels of the facet (i.e., TALQ items) are sampled from an infinitely large universe of possible levels of the facet.

4. Results

Table 1 presents results of the generalizability analysis of the TALQ using a one-facet design including students and items. As shown in Table 1, the variance component for the students accounted for 16% of the total variance, suggesting that averaging over all the items, the students in the sample differed in the assessment literacy as measured by TALQ. This variability is desirable as it reflects the systematic individual differences in the construct being measured (Alkharusi, 2012). The variance component for the items accounted for 4% of the total variance, suggesting a high level of item generalizability. This small variance component of the items suggests that the items of the TALQ appear to be consistent measuring a unitary construct of assessment literacy, theraby supporting previous psychometric analyses of the TALQ (Alkharusi, 2011a, 2011b). Items' scores can be generalized from one item to another. The residual term accounted for 80% of the total variance, suggesting that there might be other sources of variability not accounted for by differences between students, items, or both.

Table 1. Results of the generalizability analysis of the TALQ using a one-facet design

Source of variation	Sum of squares	Degrees of freedom	Mean squares	Variance component	% of total variance
Students (*s*)	378.53	258	1.47	.04	16%
Items (*i*)	94.56	34	2.78	.01	4%
Residuals (*si*,*e*)	1759.73	8772	.20	.20	80%

The TALQ could be used for making decisions about whether a teacher possess an adequate level of the assessment literacy (Alkharusi, 2011). As such, if decisions were based on the absolute values of the TALQ's observed scores, then using the results of the one-facet design presented in Table 1, the generalizability coefficient was estimated to be .87. The absolute error of variance estimate was found .006, suggesting that the aforementioned generalizability coefficient represents the optimal level of generalizability of the TALQ's scores based on the 35 items.

5. Discussion

Appropriate classroom assessment practices have been identified as one of the important component in the teaching and learning process due to their central role in maximizing student achievement and motivation (Lukin, Bandalos, Eckhout, & Mickelson, 2004; Nolen, 2011). Educators have long contend that teachers need to have an adequate level of assessment literacy (Popham, 2006). The concept of assessment literacy was first introduced by Stiggins (1991) as involving knowledge and skills of what it is being assessed, why it is assessed, how best to assess it, how to make a representative sample of the assessment, what problems can occur within the assessment process, and how to prevent them from occurring. At the same time, a set of “Standards for Teacher Competence in Educational Assessment of Students” was jointly developed by the AFI, NCME, and NEA (1990). Further, Brookhart (2011) has provided an outline of the knowledge and skills teachers should have regarding educational assessment and relates these knowledge and skills to the current teacher assessment needs in the context of formative and standards-based assessment. Several professional development programs have proposed to acquire teachers with the adequate level of assessment literacy necessary for appropriate classroom assessment practices (Koh, 2011; Lukin et al., 2004; Xu & Liu, 2009).

In response to the aforementioned concerns, Plake and Impara (1992) developed the TALQ as a means to measure teachers’ assessment literacy. Given the increasingly high-stakes nature of educational accountability, it is critical to examine the score reliability of the TALQ to make certain that it provides meaningful and dependable measures of teachers’ assessment literacy. Thus, the present study utilized GT to estimate the variance components of pre-service teachers’ scores from TALQ. The results showed that the pre-service teachers differed systematically in the assessment literacy as measured by the TALQ. Also, the analysis supported the assumption that the TALQ measures one factor of general assessment literacy (Alkharusi, 2011a, 2011b). The high level of generalizability coefficient of .87 supports the dependability of the TALQ in appraising the effectiveness of the professional development programs designed to improve teachers’ assessment literacy.

Overall, the results support the utility of using TALQ with the purpose of providing educators with information about the level of knowledge and skills possessed by the teachers. Yet, the increasing interest on the role of classroom assessment on student outcomes require more research regarding the validity and reliability of scores obtained from the available instruments measuring teachers’ assessment literacy. Research is further needed to explore how different personal characteristics, assessment background, and contextual factors influence the reliability of teacher performance assessment.

6. Conclusion

Score reliability refers to the consistency of the observed scores obtained from a particular instrument. Cronbach’s alpha is one of the most frequently reported measure of the internal consistency of the observed scores obtained through a single administration of an instrument. It provides a reliability estimate to determine if the items of the instrument are consistently measuring the hypothesized construct of interest (Crocker & Algina, 1986). However, it considers the items as the only source of the measurement error in the assessment of the score reliability (Alkharusi, 2012). The present study utilized the GT to identify the sources of error (items and students) that may affect score reliability of the TALQ. Examining score reliability of the TALQ is important due to the increasingly high-stake nature of the educational accountability as it relates to student performance on testing which rely on teachers' assessment literacy level.

The generalizability results show that the TALQ's scores are highly generalizable across items with relatively small item variance components. These results suggest that desirable levels of score reliability can be achieved with 35 items in TALQ. In this study, students represent a facet of differentiation, and as such it contribute desirable variance. The results showed a relatively large variance component for the students facet. Parallel with previous studies (Alkharusi, 2011a, 2011b), the results from the present study demonstrate the ability of the TALQ to differentiate between the various levels of assessment literacy of the teachers.

References

Alkharusi, H. (2011a). An analysis of the internal and external structure of the teacher assessment literacy questionnaire. *The International Journal of Learning, 18*, 515-528.

Alkharusi, H. (2011b). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia Social and Behavioral Sciences, 29*, 1614-1624. http://dx.doi.org/10.1016/j.sbspro.2011.11.404

Alkhrausi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education, 2*, 184-196.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9*, 30 – 32. http://dx.doi.org/10.1111/j.1745-3992.1990.tb00391.x

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*, 3-12. http://dx.doi.org/10.1111/j.1745-3992.2010.00195.x

Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the meeting of the Mid-Western Educational Research Association, Columbus, OH.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group/ Thomson Learning.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*, 255-276. http://dx.doi.org/10.1080/10476210.2011.593164

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*, 26-32. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00156.x

Nolen, S. B. (2011). The role of educational systems in the link between formative assessment and motivation. *Theory Into Practice, 50*, 319-326. http://dx.doi.org/10.1080/00405841.2011.607399

Plake, B. S., & Impara, J. C. (1992). *Teacher competencies questionnaire description*. Lincoln, NE: University of Nebraska.

Popham, W. J. (2006). Needed: A dose of assessment literacy. Educational *Leadership*, *63*, 84 – 85.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*, 238 – 245.

Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly*, *43*, 493-513.