ELSEVIER

Teaching and Teacher Education 21 (2005) 357–364

TEACHING AND TEACHER EDUCATION

www.elsevier.com/locate/tate

Item-writing rules: Collective wisdom

Bruce B. Frey*, Stephanie Petersen, Lisa M. Edwards, Jennifer Teramoto Pedrotti, Vicki Peyton

Department of Psychology and Research in Education, School of Education, University of Kansas, 1122 West Campus Road, Room 643, Lawrence, KS 66045, USA

Received in revised form 5 January 2004

Abstract

In student assessment, teachers place the greatest weight on tests they have constructed themselves and have an equally great interest in the quality of those tests. To increase the validity of teacher-made tests, many item-writing rules-of-thumb are available in the literature, but few rules have been tested experimentally. In light of the paucity of empirical studies, the validity of any given guideline might best be established by relying on experts. This study analyzed twenty classroom assessment textbooks to identify a consensus list of item-writing rules. Forty rules for which there was agreement among textbook authors are presented. The rules address four different validity concerns—potentially confusing wording or ambiguous requirements, the problem of guessing, test-taking efficiency, and controlling for testwiseness.

Keywords: Classroom assessment; Item-writing rules; Teacher-made tests

1. Introduction

Classroom assessment is an integral part of teaching (Chase, 1999; Popham, 2002; Trice, 2000; Ward & Murray-Ward, 1999) and may take more than one-third of a teacher's professional time (Stiggins, 1991), yet there are few research-based rules to guide teachers in this activity. Teachers of classroom assessment must rely on advice, opinion, experience, and common sense to direct their students in constructing classroom tests that produce reliable and valid scores. In the absence of empirical research, what rules can educational researchers provide for those who produce classroom assessments? The purpose of this study was to analyze 20 popular classroom assessment texts to identify, through group consensus, the recommended practices (or rules-of-thumb) for writing paper-and-pencil objectively scored classroom assessments. Additionally, recommended practices consistent with the few empirically based research studies that do exist were identified.

*Corresponding author. Tel.: +1 785 864 9706.
E-mail address: bfrey@ku.edu (B.B. Frey).

0742-051X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tate.2005.01.008

2. Review of the literature

Most classroom assessment involves tests that teachers have constructed themselves. It is estimated that 54 teacher-made tests are used in a typical American classroom per year (Marso & Pigge, 1988) and worldwide, millions of unique assessments, perhaps billions, are produced yearly (Worthen, Borg, & White, 1993). Regardless of the exact frequency, teachers regularly use tests they have constructed themselves (Boothroyd, McMorris, & Pruzek, 1992; Marso & Pigge, 1988; Williams, 1991). Further, studies of teachers in the United States indicate that they place more weight on their own tests in determining grades and student progress, than they do on assessments designed by others, or on other data sources (Boothroyd et al., 1992; Fennessey, 1982; Stiggins & Bridgeford, 1985; Williams, 1991). Many teachers believe that they need strong measurement skills (Wise, Lukin, & Roos, 1991), and report that they are confident in their ability to produce valid and reliable tests (Oescher & Kirby, 1990; Wise et al., 1991). Other teachers, however, report a level of discomfort with the quality of their own tests (Stiggins & Bridgeford, 1985) or believe that their training was inadequate (Wise et al., 1991). Indeed, most US state certification systems and half of all teacher education programs in the US have no assessment course requirement or even an explicit requirement that teachers have received training in assessment (Boothroyd et al., 1992; Stiggins, 1991; Trice, 2000; Wise et al., 1991). In addition, teachers have historically received little or no training or support after certification (Herman & Dorr-Bremme, 1984). The formal assessment training teachers do receive often focuses on large-scale test administration and standardized test score interpretation, rather than on the test construction strategies or item-writing rules that teachers need (Stiggins, 1991; Stiggins & Bridgeford, 1985).

A quality teacher-made test should follow valid item-writing rules, but as many researchers point out, empirical studies establishing the validity of item-writing rules are in short supply and often inconclusive, and, "item writing-rules are based primarily on common sense and the conventional wisdom of test experts" (Millman & Greene, 1993, p. 353). Even after decades of psychometric theory and research, Cronbach (1970) bemoaned the almost complete lack of scholarly attention paid to achievement test items. Twenty years after Cronbach's warning, Haladyna and Downing (1989a) reasserted this claim, stating that the body of knowledge about multiple-choice item writing was still quite limited and added recently that "item writing is still largely a creative act" (Haladyna, Downing, & Rodriguez, 2002, p. 329). The current empirical research literature for item-writing rules-of-thumb is most often of two kinds: (a) studies which look at the relationship between a given item format and either test performance or the psychometric properties of the test; and (b) studies which have evaluated the quality of teacher-made tests by applying some set of item-writing standards or criteria. Reviewing these studies for an agreed upon list of classroom assessment rules, however, is not overly fruitful, as few rules present themselves.

Haladyna and Downing (1989a, b) and Haladyna et al.(2002) have cataloged guidelines for multiple-choice, matching and alternate-choice (e.g. true-false) items with at least some evidence of validity by examining textbook endorsement and empirical studies. Though the authors did find empirical support for general advice such as "avoid trick items" and many studies testing particular rules, only four specific rules on their final revised inventory were supported without contradiction across studies and two of those were supported by the existence of only one study. It is unclear why, relative to other psychometric areas, so little research has been published. For those few studies, however, the evidence does support the particular rules. Our search of additional recent literature (1989 to present) found little beyond Haladyna et al.'s exhaustive review (2002) and focused on the same few empirically validated rules (Klein & Klein, 1998; Knowles & Welch, 1992).

Though there has, of late, been greater research emphasis on the importance and value of other types of assessments in the classroom (e.g. performance-based, *authentic*, formative, and informal), the majority of tests that teachers