Studies in Educational Evaluation 47 (2015) 12–18

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Changing the guard: Staff turnover as a source of variation in test results

Ivo J.M. Arnold *

Erasmus School of Economics, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form 24 April 2015
Accepted 25 May 2015
Available online 6 June 2015

Keywords:
Standard setting
Test difficulty
Test result variability

ABSTRACT

Due to variation in test difficulty, the use of pre-fixed cut-off scores in criterion-referenced standard setting methods may lead to variation in grades and pass rates. This paper aims to empirically investigate the strength of this relationship. To this end we examine a dataset of over 500 observations from an institution of higher education in The Netherlands over the period 2008–2013. We measure variation in test difficulty by using students' perceptions of the validity of the examination and by recording personnel changes in the primary instructor. The latter measure is based on the considerable variation in teachers' ability to assess test difficulty that is found in the literature. Other explanatory variables are course evaluations, instructor evaluations and self-reported study time. Variation in student quality is controlled for by measuring course results in deviation from the cohort average. We take a panel approach in estimating the effect of the explanatory variables on the variability in grades and pass rates. Our findings indicate that exam validity and instructor change are significantly related to variation in test results. The latter finding supports the hypothesis that instructors' difficulty in assessing test difficulty may introduce subjectivity in criterion-referenced standard setting methods.

1. Introduction

The vast literature on standard setting in assessment testifies to the importance of high quality testing in education. Both students and society at large expect assessments to be valid and reliable measurements of academic performance. An assessment is valid when it "reflects the material covered in the educational program, taking the academic standard of the students into account" (van de Watering & van de Rijt, 2006, p. 134). Notwithstanding the importance of testing and standard setting, the literature offers no consensus on what the "best" standard setting method is (Downing, Tekian, & Yudkowsky, 2006; Friedman Ben-David, 2000).

There are two basic types of standard setting methods. Using relative or norm-referenced methods, the cutpoint is set as a percentage of examinees, implying that standard setting depends on test results (Norcini, 2003). In contrast, when absolute or criterion-referenced methods are used, standards are set as a percentage of test questions. As a result, in the latter case standard setting is independent of test results. In general, relative methods are considered more appropriate when the purpose of the test is to rank examinees or to identify or select a specific number of examinees. Absolute methods are preferred when the purpose is to establish whether examinees meet pre-set requirements (Norcini, 2003).

Absolute methods ideally require the use of expert panels to predict how examinees will score on test items. For example, the Angoff (1971) method requires panelists to predict the proportion of test items answered correctly by minimally competent candidates. For frequent testing within a large educational institution, the use of expert panels is costly and time-consuming. In practice, therefore, absolute standard setting is frequently applied using pre-fixed cutpoints. A potential drawback of using criterion referenced standard setting with pre-fixed cutpoints is that test difficulty is insufficiently taken into account. Fluctuations in test difficulty may then induce undesirable variation in grades and pass rates. Norm-referenced methods also have drawbacks, such as their tendency to weaken the relationship between examinees' knowledge and abilities and the pass-fail decision. In addition to these two basic types of standard setting, a number of compromise methods has been developed. Examples are the Hofstee (1983) method and the Cohen method (Cohen-Schotanus & Van der Vleuten, 2010), which both aim to correct for test difficulty. Norcini (2003) provides a comprehensive overview.

The claim that criterion-referenced standard setting with pre-fixed cutpoints does not account for variation in test difficulty

* Tel.: +31 104081254.
E-mail address: arnold@ese.eur.nl

http://dx.doi.org/10.1016/j.stueduc.2015.05.002
0191-491X/

derives from the rigid nature of the cutpoints. With pre-fixed cutpoints, any variation in test difficulty will logically translate into variation in grades and pass rates. Empirical evidence on whether this theoretical problem is also of practical relevance is, however, limited. This paper therefore aims to empirically investigate the effect of test difficulty on the variation in test results when criterion-referenced standard setting using pre-fixed cut-off scores is used. We do this by relating measures of test difficulty to variation in pass rates and grades, controlling for variables related to course, instructor and student quality.

We start with two observations that can be made from the existing literature. First, criterion-referenced methods may lead to high variability in test results. For example, Cohen-Schotanus and Van der Vleuten (2010) report that failure rates of 52 medical tests at Groningen University varied from 17% to 97% using pre-fixed cut-off scores. The authors conclude that this "...seems to point to a major effect of variation in test difficulty" (p. 156). Second, a large literature documents that teachers are in general unable to correctly estimate the difficulty levels of assessment items (Clauser, Clauser, & Hambleton, 2014; Goodwin, 1999; Impara & Plake, 1998; Plake & Impara, 2001; Shepard, 1995). The review by van de Watering and van de Rijt (2006) indicates that teachers tend to underestimate the level of difficult items and overestimate the level of easy items. They conclude that: "In higher education, results show that teachers are able to estimate the difficulty levels correctly for only a small proportion of the assessment items" (p. 133). A possible explanation for this bias is that, as experts in their field, teachers have difficulty to project themselves into the position of students (Goodwin, 1999). Clauser et al. (2014) also find that there is considerable variation in the ability of experts to estimate item difficulty consistently. In his review of studies related to the Angoff method, Brandon (2004) concludes that the method lacks the desired validity, as measured by the deviation of experts' estimates from empirical item p-values. However, the literature also shows that training, information sharing and group discussions among expert judges increase the validity (Plake & Impara, 2001; Brandon, 2004).

From these two observations, one may infer that the variability in test results using criterion-referenced standard setting is (partly) due to teachers' inability to correctly assess test difficulty. This view is also commonly held by researchers in this field. For example, Cohen-Schotanus and Van der Vleuten (2010, p. 156) postulate that the "most probable cause" for variability in pass/failure rates is the variation in test difficulty. Yet empirical evidence firmly linking the variability in test results to measures of test difficulty is, to our knowledge, lacking. This lack of evidence is an omission, as the view of these researchers is at odds with explanations commonly favored by teachers themselves. Cohen-Schotanus and Van der Vleuten (2010) report the following common explanations for high failure rates: "students do not study hard enough; class attendance is low; the previous cohort was much more intelligent or students were preoccupied with the Soccer World Championship in the run-up to the test" (p. 157). Such anecdotic evidence may sound familiar to professionals working in education management. But in the absence of empirical evidence on the sources of variability in failure rates, these explanations go unchallenged.

Teachers' explanations for high failure rates are compatible with a self-serving bias, which occurs when people attribute success to internal factors but failure to external factors beyond their control (Miller & Ross, 1975). In the current context, the external attribution of the cause of high failure rates may result from teachers' need to protect their self-esteem following their failure to make an adequate exam or their need to cope with the disappointing test results. In the absence of empirical evidence linking test difficulty to test results, the self-serving bias hypothesis remains speculative.

This paper aims to estimate the relationship between the variation in test results and measures of test difficulty for exams using criterion-referenced standard setting with pre-fixed cut-off scores. We hypothesize that measures of test difficulty are related to the variation in test results. One measure that we use is students' perception of the validity of the examination. An assessment which in students' eyes is not valid will be regarded as more difficult. Assessment validity requires that an exam reflects the material covered in the course (van de Watering & van de Rijt, 2006). Under the plausible assumption that exams that do not reflect the course content are harder to pass, this is our first measure of test difficulty. The use of student perceptions is supported by evidence that students are better predictors of test difficulty than teachers (Verhoeven, Verwijnen, Muijtjens, Scherpbier, & Van der Vleuten, 2002). In addition we examine whether turnover in the staffing of courses can explain part of the variability in test results. This choice follows logically from the finding in the literature that individual teachers have difficulty in predicting test difficulty, while group discussions typically reduce the prediction error (Brandon, 2004). If instructors hold subjective views on test difficulty, a changing of the guard may affect test results. Needless to say, if changes of the guard are an important source of variation in test results, this would seriously question the credibility of assessment.

In addition to measures of test difficulty we include a number of control variables related to the quality of the course, the instructors and the student. Most of these variables derive from the student evaluation of teaching (SET). The use of SET scores to measure and evaluate the quality of teaching is highly controversial in higher education (Theall & Franklin, 2001). There is also no consensus on the strength and the interpretation of the relationship between SET scores and course grades (Gump, 2007). The present paper does not aim to provide new evidence on the validity of SET scores or their relationship with course grades. However, as empirical studies often find a positive relationship between SET scores and course grades, see e.g. Brockx, Spooren, and Mortelmans (2011), we feel the need to include SET scores as intervening variables, to avoid estimating a spurious relationship between measures of test difficulty and variation in test results.

This paper aims to make the following contributions to the literature. First, while the available evidence on teachers' difficulty in correctly assessing test difficulty can serve at most as circumstantial evidence supporting a link between test difficulty and test results, our approach allows for a direct estimation of this relationship. Second, one of our empirical measures of variation in test difficulty – instructor change – is, to the best of our knowledge, novel. Measuring the direct impact of staff turnover on the variation in grades and pass rates has not been done before.

Our empirical approach is to estimate the relationship between variation in test results and test difficulty using a panel regression model which controls for a number of intervening variables. Before proceeding to the research methodology, the next section first describes the setting of this study and the nature of the dataset.

2. Setting and data

This study was conducted at a School of Economics which is part of one of the research universities in the Netherlands. The school offers three bachelor programs in the economics discipline. The programs share a common educational system and design. The nominal duration of the programs is three years. The first two bachelor years consist of obligatory core and support courses. Most courses are delivered using a combination of large-scale plenary lectures and small-scale tutorials with required attendance. During the period of investigation, no major changes in the educational system have taken place. Regarding the curriculum, minor changes in the line-up of courses have been implemented.