Cut scores for pass/fail decisions

9.1 Introduction

Many educational and licencing tests are used for classifying examinees into a small number of categories, such as ‘satisfactory’, ‘borderline’, and ‘unsatisfactory’. In such tests, each possible response pattern has to be assigned to one of the categories. This task is often simplified by defining a scoring rule for the responses, so that the examinees are ordered on a unidimensional scale. Then it suffices to set the cut scores that separate the categories. When several test forms are used, separate sets of cut scores have to be set for each of them because the forms are likely to have different score distributions.

Without substantial loss of generality, we consider the case of a single cut point separating two categories of examinees. For instance, a test may be used to assign examinees to a ‘pass’ and a ‘fail’ category, and the cut score can be interpreted as the minimal acceptable competence. We refer to the examinees with scores above the cut score as competent. It is important to distinguish between observed-score and true-score competence because of the two kinds of incorrect decisions that can be made about each examinee: failing a true-score competent examinee and passing an examinee whose true-score is below the cut point.

If we knew that a given percentage of the examinees were competent, or simply wished to pass a certain percentage of the examinees, we could assign that percentage of the examinees with the highest observed scores to the ‘pass’ and the remainder to the ‘fail’ category. If there are costs
associated with each incorrect decision, this assignment scheme has to be adjusted. For example, if passing an examinee who should have been failed is associated with extremely high costs, only examinees who have performed very well should be passed.

In a more realistic setting, experts are engaged to set the cut score for passing. Of course, the experts could classify every examinee by his/her responses. This is not a workable proposition in large-scale tests because a large number of experts would be required. Also, the objectivity of the classification process would be undermined, unless the examinees were classified by several experts. Procedures for resolving disagreement among the experts assessing the same examinee would have to be introduced. The adjustment schemes discussed in Chapter 3 provide one avenue for doing this.

The objectivity of the classification process remains intact if the experts are engaged merely to set the rules for classification. There are two distinct approaches. In the first approach, each expert is assumed to have a good idea about the kind of examinee who is on the borderline between ‘fail’ and ‘pass’. For each item of the test form, the expert gives his/her view of the chance that such a borderline examinee would give the correct response. For items such as essays, scored on an ordinal scale, the expert guesses the average grade for such examinees. The cut score is set to the expected test score for a borderline examinee. The experts’ expectations are averaged in the obvious fashion. Such schemes have been extensively explored by Nedelsky (1954) and Subkoviak (1978); see Angoff (1971) and Hambleton et al. (1978) for reviews.

The principal drawback of such schemes is that the experts are often very inconsistent in their perception of percentages or fractions. Their task can be made easier by asking them to classify each item into one of an ordered set of categories, such as ‘almost nobody’, ‘some’, ‘about half’, and ‘almost everybody’, according to the proportion of borderline examinees who they believe would give the correct response. The analyst’s task is to attach ‘typical’ percentages to these categories. In this way, some consistency in the experts’ assessments may be introduced at the price of a coarse classification.

The subject of this chapter is another scheme for eliciting and reconciling experts’ proposals for cut scores. In this scheme, the experts inspect the sets of responses of a small number of examinees and classify each examinee as ‘fail’ or ‘pass’. Independently, the responses of each of these examinees are scored. The experts’ classifications are then related to the scores and the (fitted) association is applied to all the examinees, including those whose response sheets were inspected by the experts. In this way, the experts are engaged only for a limited time and the classification procedure is objective. Note the pivotal role played by the test score.

The topic of this chapter is closely related to that of Chapters 2–4. We are interested in reconciling experts’ ratings and, as a form of quality control,