AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY

AMERICAN GUIDANCE SYSTEM

AMERICAN GUIDANCE SOCIETY
Norm-referenced vs. criterion-referenced standards

In credentialling examinations where the stakes are high, standard-setting approaches may differ in their orientation of norm referenced vs. criterion referenced. Credentialling organizations may use norm-referenced orientation, in which the standard is based on performance of an external large representative sample (norm group) equivalent to the candidates taking the test. The criterion-referenced orientation links the standard to the content of the competence level under consideration. The norm-referenced standard will be somewhat unstable and will shift according to the performance of the norm group, as large as it may be. The criterion-referenced standard is a fixed standard that may undergo periodic re-evaluation in view of shifts or trends in candidates’ performance over time (Nungester et al., 1991). The norm-referenced approach, employing a group-referenced standard, may result in reasonable standards providing the group is representative of the candidates’ population, is heterogeneous and is large. Shift of the standard over time is a concern.

At the school level, relative (norm) vs. absolute (criterion) standards are also considered (Muijtjens et al., 1998). A relative standard can be set at the mean performance of the candidates, or by defining the units of standard deviation from the mean. Percentile ranks, as well as the median point, could also be set as norm-referenced standards. These standards may vary from year to year due to shifts in the ability of the group and may result in a fixed annual percentage of failing students, if the scores maintain a normal distribution across administrations. An absolute standard stays the same over multiple administrations relative to the content specifications of the test. The failure rate may vary due to changes in the group’s ability, from one administration to the other.

Nedelsky stated:

The passing score is to be based on the instructor’s judgement of what constitutes an adequate achievement on the part of the student and not on the performance by the student relative to his/her class or to any other particular group of students. In that sense the standard to be used for determining the passing score is absolute. (Nedelsky, 1954, p. 3)

A standard is defined as absolute “if it can be stated in terms of the knowledge and skills a student must possess in order to pass the course” (Nedelsky, 1954, p.4).

Compensatory vs. conjunctive standards

An example of an objective structured clinical examination (OSCE) will help clarify the difference between compensatory and conjunctive standards. Let us assume we are required to set standards on a 15 stations standardized patients examination. Each station measures three skills: history, physical examination and communication. Each station contains separate checklist items (criteria) for each skill component. Let us further assume that each station varies significantly in difficulty in relation to history and physical examination scores. However, performance on communication skills does not vary substantially from one station to another, and a constant difficulty level is maintained for all stations.

In any given station, all candidates’ positive responses per skill are summarized and divided by the highest possible score and multiplied by 100 to produce a station skill percentage score. The total test score is calculated by averaging the station scores across the 15 stations. If the standard is set on the total test score, which is the average performance across stations, the skill standard will constitute a compensatory standard. This method of scoring permits candidates to compensate for relatively poor skill performance on some stations (Norcini et al., 1993). Any combination of performance (skill scores) across the stations is acceptable, as long as the examinee exceeds the skill performance standard for the total test (Hambleton, 1995).

Test developers or decision makers may, however, insist that a number of stations should be competently managed before a passing score is warranted. This method, which explicitly defines the number of stations that must be successfully managed in order to pass the test, constitutes a conjunctive standard. Similarly, test developers may decide that candidates must exceed separately each skill standard (history, physical examination and communication) at the total test level. This approach will also constitute a conjunctive standard, which does not allow candidates to compensate for relatively poor performance on one or more skills. For example, poor performance on history should not be compensated by high performance on physical examination and vice versa. However, multiple conjunctive standards may produce multiple failures. In a long OSCE examination, even the better students may fail a station because of measurement error (Hambleton, 1995). In an OSCE a single station score is considered an unreliable score, mostly as a result of case specificity and sampling error (C.P.M. van der Vleuten, G.R. Norman & E. De Graaf, Pitfalls in the Pursuit of Objectivity, personal communication, 1990). Acceptable test reliabilities are attained with careful choice of optimal number of stations, which will allow generalization to other clinical cases. Standards should be set on reliable scores to avoid problems of decisions inconsistencies and candidates’ misclassifications (Jaeger et al., 1996). Thus, conjunctive standards set on scores with low reliability at the station or test level may result in wrong pass/fail decisions due to measurement error. Issues of retesting the examinations due to multiple failures and the logistics and cost involved should be considered carefully, prior to selecting a conjunctive approach. The degree of relationship among the test skill components may further assist in conjunctive vs. compensatory decisions. The higher the correlation among the test components the greater the inclinations towards a compensatory standard. Highly correlated test components may imply a common construct or dimension of performance, in which performance on one component impacts on performance on the other. The

Weaknesses of relative/norm-referenced standards

- Standards are not content related
- A fixed number of candidates may fail each year
- Examinees’ ability may influence the standard
- Standard is not known in advance
- Diagnostic feedback relative to performance is unclear
aggregation of related component scores into one dimension score is the basis for a compensatory pass/fail standard. If there is no or little relationship among test components, the conjunctive methods could be justified. The communication skill in this example may call for a compensatory standard to be set for the total test and not separately for each station since skill scores do not vary significantly from station to station. However, the history and physical examination standards will be set separately for each station and it is the decision of the test developers whether the total test standard for each skill will be conjunctive or compensatory. For example, in an OSCE where history and physical examination test component scores are highly correlated, both components may be regarded as one dimension of data-gathering skills. Setting standards on a combined score of data gathering constitutes a compensatory approach by which an examinee may compensate for poor physical examination skills with good history skills. In another example the two skills components may be uncorrelated but a combined score presents higher reliability compared with the two unreliable, separate component scores. In such a case psychometricians will explore issues of misclassification in the aggregated approach (compensatory) vs. the separated component approach (conjunctive) (Jaeger et al., 1996). A conjunctive standard allows diagnostic feedback to candidates, since each skill component is considered separately. In the compensatory approach information about test components is aggregated by setting one standard. Consequently, performance feedback on separate test components may not be available.

Standard-setting methods

Conjunctive standard considerations

- Reliability of the test components at the station or skills level might be low
- Long tests may produce multiple failures at the station level
- Standards on multiple-test components may also result in multiple failures

Compensatory standard considerations

- What are the consequences for compensating for poor performance?
- Loss of diagnostic feedback

Test-centred models

Jaeger (1989) refers to test-centred models in which the judges set standards by reviewing the test items and provide judgements as to the ‘just adequate’ level of performance on these items. The following descriptions of the test-centred method accentuate the different ways for judges to provide their ratings. The Angoff (1971) model employs a test-centred approach and is known for its wide use in educational testing and performance assessment. The present guide will use the Angoff approach as a case study for setting standards.

Ebel’s (1972) approach requests judges to categorize the items in a test employing a number of categories, for example: three levels of difficulty and four levels of relevance to the decision to be made. In a modified Ebel procedure, judges categorize items as ‘Essential’, ‘Important’ or ‘Indicated’ (Case & Swansen, 1998). After classifying the items into each category, judges then decide on the proportion of items in each category that a hypothetical group of borderline examinees could respond to correctly. The number of items in each category is multiplied by the proportion of items in that category that would be answered correctly by a borderline candidate, and these products are summed to produce a test standard.

The Nedelsky (1957) approach was originally designed for multiple-choice items. For each item, the judges decide on how many of the distractors (response options) a minimally competent examinee would recognize as being incorrect.

The minimal pass level for the item is computed as one over the number of distractors (options) remaining after the incorrect options are removed. For example, on a five-option item, if a judge decides that the minimally competent examinee will recognize two options as being incorrect, the minimal pass level (MPL) for this item will be 0.33, since three options remain after the two incorrect options were removed. The MPL is then summed over all items in the test to produce a passing score for the total test.

Jaeger’s method (1982, 1989) differs from most other test-centred methods in a number of important ways (Kane, 1994), all of which tend to emphasize the role of standard setting as the development of policy rather than as a technical approach to estimating a passing score. Jaeger (1989) emphasizes the importance of recognizing the need to “sample all populations that have a legitimate interest in the outcomes of competency testing”. The political aspects involved in many testing contexts is emphasized as well as the need to incorporate multiple points of view. Most of the other methods employ a single panel of expert judges. The focus in Jaeger’s method is on the passing examinees rather than on the borderline or the minimally competent. Judges are asked to rate ‘yes’ or ‘no’ for each item, whether each examinee who passes the test should be able to answer the item correctly. This procedure is employed within a number of expert panel groups. Each panel will receive input from the other panel’s ratings and will be presented with actual performance data. Judges are then asked to provide second ratings or more, while reconsidering their judgements in view of the new information. In such iterative procedures, considerations of the validity of the standards are built into the process (Kane, 1994).

Examinee-centred models

In the examinee-centred approaches panelists make pass/fail decisions by identifying a point on the score scale that would be most consistent with the test purpose (Kane, 1994).

In the Borderline-Group method (Livingston & Zieky, 1982), the judges may use different approaches to identify an actual (not hypothetical) borderline group. The level of achievement of this group is around the performance standard. The judges may use their experience or other methods of assessment to identify the group. The median score for this group could be used as the passing score. The
scores of the borderline group should cluster together to produce a reasonable standard. If the scores are spread, the method may not be applicable.

In the Contrasts by Group approach (Livingston & Zicky, 1982), the panellists sort the examinees into two groups: competent and not competent. The judgement is based on characteristics of the examinees relative to the task other than the test scores (i.e., the test scores are not known to the panellist during the sorting process). After the sorting is completed, the score distributions for the competent and not competent groups are plotted. Commonly, the point of intersection of the two distributions could be considered as the passing score (Clauser et al., 1996; Burrows et al., 1999). However, the overlapping areas of the distribution would indicate the degree of separation between the two groups. The larger the overlapping areas, the smaller the separation between the two groups. Ideally the chosen passing score should maximize discrimination between the competent and the non-competent examinees, thus the distributions should indicate minimal overlap.

Relative/absolute compromise standards: the Hofstee method
This is a standard-setting approach that incorporates the advantages of both relative and absolute standard setting procedures (de Gruijter, 1985). Panellists review the test materials and are asked to provide four values:

1. lowest acceptable percentage of failing examinee (minimum failure rate);
2. highest acceptable percentage of failing examinee (maximum failure rate);
3. lowest score which allows a candidate to pass (minimum passing point);
4. highest score required for a candidate to pass (maximum passing point).

The median values of the group of judges are plotted for each value.

In Figure 1, the minimum failure rate is 0, the maximum failure rate is 20%, the minimum passing point is 50%, and the maximum passing point is 60%. The (actual performance) test scores curve is plotted, which indicates the failure rate as a function of the passing score. The intersection of test scores curve with the diagonal line drawn from upper left to lower right is the cut-off point (just above 55% correct).

Modified Angoff
The Angoff standard-setting approach is one of the most widely used in medicine (Cizek, 1996). In its purest form the Angoff method is a judgemental approach in which a group of expert judges makes estimates about how borderline candidates would perform on items in the examination (Livingston & Zicky, 1982; Berk, 1986), i.e., the proportion of borderline examinees who will answer an item correctly. This is equivalent to estimating the candidate’s likelihood of answering a number of items correctly (Sizmur, 1997). Estimates are averaged over judges and summed over items to create a standard (cut-off score). A modification adds to the process by the provision of data about difficulty of the items, based on actual performance data. The focus on the borderline group should enhance maximum discrimination at the borderline level.

The Angoff method depends on expert judges’ (panellists’) degree of familiarity with the minimally competent test-taker (the hypothetically borderline group), who is neither qualified nor unqualified to pass the test (Norcini et al., 1993). The panellists are asked to make judgements about that candidate’s likelihood to respond correctly to each of the test items. This task is highly dependent on the capacity of the panellists to understand the meaning of a borderline candidate’s characteristics. The panel list also has to understand how the borderline candidate will respond to the required performance task. The panelist must also be familiar with the expected level of examinee performance as defined by the purpose of the test, and be able to understand the degree of difficulty of the task. Research with judges showed that the standards that they demanded in theory were extremely stringent and different from those they actually used in practice. In general, judges have a tendency to produce very high standards (Wolf & Silver, 1986).

This guide for standard-setting approaches will focus on one modified version of the Angoff approach. In reviewing the scoring sheet of an OSCE station, panellists will be asked to determine the number of scoring points an individual borderline candidate will answer correctly in order to pass the station. This process will be done separately for each of the station’s skill components: history and physical examination. A separate process will be outlined for communication skills.

<table>
<thead>
<tr>
<th>Modified Angoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Test-centred approach</td>
</tr>
<tr>
<td>• Criterion referenced</td>
</tr>
<tr>
<td>• Relies on expert (panellist) judgements</td>
</tr>
<tr>
<td>• Requires understanding of borderline characteristics</td>
</tr>
<tr>
<td>• Actual performance data are provided to panellists as an additional information source</td>
</tr>
</tbody>
</table>

Selection of panellists
The importance of panellist selection for the creation of a defensible standard-setting process cannot be overemphasized. Hambleton (1995) describes a defensible process which is based on: selection of appropriate panellists, excellent panellist training and a well-planned and systematic standard-setting process, which provides ample
opportunity for discussion and deliberations among panelists. The reliability of the performance standard will be reflected by the degree of agreement among panellists and a reasonable performance standard will coincide with expectations about the specified professional performance.

Messick (1989, p. 15) refers to test validity as the “inferences and actions based on test scores”. The standard-setting procedures are in fact enhancing the validity of the test by establishing, through a defensible process, the meaning and inferences that we make on the basis of test scores (Wiliam, 1996). Thus, standard-setting procedures lie at the very heart of the validity of the test. A defensible approach to standard-setting procedures is essential to support the appropriateness of standard inferences and the consequential decisions based on those inferences.

Jaeger (1991) identifies the qualifications of panelists. Among other things, they should be experts excelling in their specialization, able to conceptualize and problem-solve in their specialty and able to employ self-monitoring skills. Jaeger’s criteria for panelist selection may further be developed to include qualifications relevant to test specifications, purpose of the test and standard-setting methods employed. For example, in setting standards for a 15 station OSCE examination for fourth-year medical students who are required to pass the examination for graduation qualifications, additional panelist characteristics should be considered. Panellists must be familiar with performance of fourth-year students, should be familiar with the OSCE method of assessment, should have experience as clinical teachers and should have knowledge of the next phase of training the graduates are expected to enter.

The developers of tests must create their own list of panellists’ characteristics, which will specifically address the purpose of the examination. In a high-stakes examination, panelists who are recognized for contribution in the field may be appropriate. In addition, an adequate mix of gender, age, seniority, academic vs. community experiences and other relevant attributes will further establish the defensibility of the standard-setting process. Multiple panellist groups for rating the same stations will further add to the defensibility of the process by demonstrating consistency of standards across occasions and across panelists. The larger the panelist sample the more stable will be the resulting standard (Jaeger et al., 1996). However, one needs to consider how manageable a large size would be. Jaeger et al. (1996) propose an estimation method for panellist group size by considering the standard error of measurement of the assessment tool, and the standard deviation of the panellist ratings.

Since multiple-choice examinations are the most commonly used form of written tests, it may be helpful to consider differences between multiple-choice examinations (MCQ) as a form of written test and an OSCE as a form of performance test. Let us assume that the MCQ examination contains 400 items, sampled across domains. The OSCE test contains 15 standardized patient-based stations sampled across specialty areas. Items on both forms of assessment are either a multiple-choice question or an OSCE checklist criterion. The following statements highlight the differences between the two forms of assessment and questions are raised for further consideration with regard to the modified Angoff approach.

(1) The traditional MCQ items are written independently of each other. In contrast, a station item is dependent on the other items in the checklist. Thus, panellist ratings for MCQs are provided separately for each item, whereas in the OSCE example items are either rated independently or as ‘sets’.

(2) While estimating the difficulty of an MCQ item, panelists consider the likelihood of candidates getting an isolated knowledge component right. In the OSCE example the estimation is applied to a ‘set’ of items which represent a ‘dimension’ of professional competence.

(3) The large number of MCQ items (400) challenges the generalizability of the number of OSCE stations and the broad nature of the tests. The complexity of performance tasks, in terms of resources and time, limits the test to a manageable size.

(4) In multiple-choice questions the common response is one correct answer. In an OSCE station the response may represent different degrees of success on a scale continuum.

The above are only a few differences between MCQs and OSCE items. Since traditional standard-setting methods

---

**Panellists should be:**

- Experts in the related field of examination
- Familiar with the examination methods
- Good problem solvers
- Familiar with level of candidates
- Interested in education (teachers)

**Written vs. performance standards**

Most of the standard-setting procedures employed for performance assessment were first applied to written tests.
are applied to performance assessment, the differences should be considered for the establishment of a valid standard.

**Educational benefits of standard setting**

**Faculty development**

Standard-setting procedures could be employed as a form of faculty development. The use of orientation materials such as videos, test forms, performance data etc. allows faculty to experience first hand information on candidates’ performance on the task. In some instances, faculty may observe for the first time videoed performance of students, containing criteria for acceptable performance. They may consider the information as feedback for their own clinical teaching. They may compare poor and excellent candidates’ performance to their expectations. Some may not ‘believe’ that the poor performer is actually enrolled in their medical school. The standard-setting process provides a reality check, which is not available otherwise owing to the low incidence of direct clinical observation in medical schools. They may identify the existence of persisting weaknesses, for example bedside manner. It is not uncommon to hear faculty say, “I am going back to the department to change the way I teach bedside manner”.

**Quality control of test materials**

The process of exposing faculty to test materials, scoring policy and profiles of scored performance constitutes a scrutinized quality-control procedure. Panellists in the process of reviewing test materials identify inappropriate items, which are either ambiguous or irrelevant. They may find tasks for which the criteria are incomplete. As an external group of experts, they will compare video performance with the actual scoring results, and will comment on discrepancies. Satisfying results of standard-setting procedures suggest that test developers should submit test materials for which they have 100% confidence in its quality and its scoring mechanism. Otherwise, they may lose panellists’ trust in the test. In actual standard-setting procedures panellists are asked to refrain from comments on the design of the station or the scoring policy. However, mistakes and inconsistencies, if present, create serious complications in the process of setting standards.

**Practical steps of the Modified Angoff Approach**

**Step 1: General orientation**

The facilitator presents to panellists a general overview of the OSCE test, the duration of each station, test components, scoring methods and any other information from which the panellists may benefit. Central to the orientation session is a statement regarding the purpose of the test. In the OSCE example, the purpose may read as follows:

The graduates of this medical school should demonstrate minimal competence in history, physical examination and doctor/patient communication, for the purpose of entering a supervised Pre-Registration House Officer (PRHO) training programme in the UK!

The key words are ‘minimal competence’ and ‘supervised programme in the UK’. This statement of purpose will guide panellists in their estimates as to what is expected in the UK concerning PRHO entry level skills, as well as what is expected of the medical school graduates at their exit point. The minimally competent concept and the continuing supervised training may, for example, direct panellists to consider conjunctive vs. compensatory standards. Knowing that candidates are expected to continue with their training in a supervised mode, and that poor performance in some areas will be further developed in the next phase of training, allows a compensatory approach. Alternatively they may consider that basic minimal competence in each component is essential for better future practice and thus a conjunctive approach is warranted.

In some instances, where faculty is not familiar with the OSCE stations, a mini OSCE is set as part of the standard-setting orientation procedure. Faculty may play the role of examinees, or examiners, by observing each other. This is done to avoid overestimation or underestimation of the station difficulty. Their own level of performance as experts may serve as a ‘ceiling’ effect for the standard-setting ratings. However, panellists should be cautioned that as experts they might perform short cuts, which will result in a lower score for the station. The scores they achieved in the mini OSCE could serve as good points of discussions regarding the panellists’ level of performance compared with the expected level of the candidate’s performance.

**Step 2: Orientation to a ‘practice’ station**

An example of the practice station is shown below, and the corresponding checklists are shown in Figures 2 and 3. The practice station is divided into two stations, history and physical. In each station an examiner observes the candidate and rates his/her performance on the checklist items.

**Instructions for Candidates**

**History taking**

Mr/Mrs J is a patient who complains of easy bruising. You have four and one half minutes to take relevant history from her. You will be asked to present a summary of your findings at the next station, including a working diagnosis.

**Physical examination**

At this station there is an observer and a patient. The patient has arterial peripheral vascular disease. Carry out an examination of the circulation of the lower extremities of the patient. While doing this, please explain to the observer what you are doing.

Test developers present the ‘practice’ stations to panellists. The practice stations materials are discussed. At the end of the presentation, panellists are asked to make judgements as to how many items should be answered correctly by the borderline candidate in order to pass the ‘practice’ stations.
The ‘practice’ orientation materials may include:

(1) a full description of the stations;
(2) history and physical examination checklists (Figures 2 and 3);
(3) videotapes of one low performer and one high performer for the practice stations;
(4) the panellist will have a blank checklist for the two component skills while viewing the video;
(5) the actual skill score will be presented to the panelists following the completion of each video performance.

The low vs. the high performers assist the panellist to form a range of possible performance profiles for the practice stations. The orientation may contain other components according to the nature of the test. A careful process of orientation may assist the panellists in arriving at ‘realistic’

**Examiner’s checklist**

Please ask the student for an identity label to place on the checklist

Using the marking schedule given below, allocate a mark by (−) in the appropriate column to the right of the sheet as carried out by the candidate.

0+ Not done / ½ = Done poorly/1 = Done satisfactorily

<table>
<thead>
<tr>
<th>Item</th>
<th>0</th>
<th>½</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation to patient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position patient on couch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspection for evidence of skin changes—pallor</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Inspection for evidence of skin changes—hair loss, dryness etc</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Palpation for temperature changes</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Palpation of pulses</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Femoral</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Popliteal</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Posterior tibial</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Dorsalis pedis</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Capillary refill (time)</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
</tbody>
</table>

Max 10 marks

**Figure 2. Practice station: History checklist.**

**Examiners’ checklist**

Please ask the student for an identity label to place on the checklist

Students have been given the following instructions

Mr/Mrs J is a patient who complains of easy bruising. You have **four and a half minutes** to take a relevant history. You will be asked to present a summary of your findings at the next station, including a working diagnosis.

Please ask the student to summarize the history he/she has taken at the last station.

*Note:* prompting halves the marks.

Use the marking schedule given below and tick the item in the appropriate box.

<table>
<thead>
<tr>
<th>Item</th>
<th>Not done</th>
<th>Prompted</th>
<th>Not prompted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name and age of patient</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Duration of bruising</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Site of bruising</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Precipitating factors (trauma/spontaneous)</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Other bleeding sites (guts, urine, joints, menorrhagia)</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Severity of bleeding</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Previous operations</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Previous dental extraction’</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Family history</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Drugs aspirin</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Anticoagulants</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>½</td>
<td>1</td>
</tr>
<tr>
<td>Sequence</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Working diagnosis</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Figure 3. Practice station: Physical examination checklist.**
judgements, and further facilitates a defensible standard-setting process.

**Step 3: Characteristics of borderline candidates**

The importance of understanding the characteristics of borderline candidates has already been emphasized. In the modified Angoff approach the borderline candidates constitute a hypothetical group. Therefore, it is essential to introduce a process by which the panellists will increase their understanding of the borderline characteristics.

This process becomes more meaningful following the orientation steps in which panellists viewed performance profiles as they understand the purpose of the test and are familiar with the test materials. During this step, a facilitator poses the question to the panellists:

Please indicate the characteristics of borderline candidates that are relevant to the skills measured in this test. These are fourth-year medical students who are neither unqualified nor qualified to pass the test. These candidates' scores lie around the estimated cut-off point. It is important that we characterize the group in order to estimate and make judgements about their ability to demonstrate minimal competence in the practice station.

Panellists may then write independently the characteristics of the borderline group per skill component. The panellists' statements are then posted and the facilitator discusses with the panellists each statement. Arguments and disagreements are clarified and the group reaches a consensus as to what would be an appropriate list of borderline characteristics per skill component. This process is essential to facilitate further the estimates and judgements of the panellists and to support the defensibility of the process.

**Step 4: Panellists provide ratings**

A rating form is distributed to panellists (Figure 4). The rating form will be used for the entire standard-setting procedure. Two separate forms will be distributed, namely (1) history and (2) physical examination. Each form will specify in the left-hand column the alphabetical order of the stations to be rated. The next column will outline the maximum number of points available for the station. The next two columns are provided for the first and second ratings.

In the first ratings, panellists are asked to enter their individual judgement as to how many items will be answered correctly by a borderline examinee in order to pass the history skill component in this station. (Note: it is recommended that the number of items will constitute the ratings and not percentages of items. Experience has shown that panellists may be influenced by traditional concepts of acceptable standards expressed in percentages [for example, 70% vs. 50%]. Thus, it is much 'safer' to count the items, and provide raw numbers as estimates.) Following the completion of the first rating, the facilitator presents all panellists' ratings on the board by assigning a number to each panellist. The panellists discuss their ratings. The facilitator encourages panellists with the highest and lowest ratings to reflect on their judgements. The facilitator will average the ratings of the panellists to produce a cut-off raw score for history in the practice station.

**Provision of actual performance data**

A history score distribution for the practice station is presented to panellists (Figure 5). The score distribution is generated from previous administrations of the station to a similar cohort of fourth-year medical students. The characteristics of the cohort such as gender, year, purpose of administration etc. should be presented to the panellists. Knowledge of the similarities or differences between the cohort and the potential candidates allows panellists to evaluate the appropriateness of the actual performance data. The distribution outlines the cumulative numbers and

<table>
<thead>
<tr>
<th>Station</th>
<th>Max. pts</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Practice)</td>
<td>20</td>
<td>1st ratings</td>
</tr>
<tr>
<td>B</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 4. Panellist rating form.*
percentages of students who got one item correct, two items correct, three items correct etc. The facilitator indicates the percentages of students who might fail the history skill component of the station if the panelist average ratings are applied to the distribution as a cut-off score. A discussion should revolve around the issue of the ‘consequential data’, i.e. percentage failure. Are panelists surprised? Should they expect a lower or a higher number of failures? Does the percentage failure seem to coincide with their experience? The purpose of the actual performance data is to orient the panelists to ‘realistic’ normative information. This provides another source of information, which helps panelists adjust their ratings, when they are asked to attempt a second rating on the form (see Figure 4).

As stated earlier, the higher the agreement among panelists, the more reliable is the standard. Therefore, it is hoped that in the second ratings, the panelists will adjust the scores in view of their peers’ ratings and the actual performance data. It is also hoped that the standard deviation of the ratings will decrease at the second ratings. Panelists provide their second ratings, which are posted on the board by the facilitator. A final cut-off score is calculated by averaging all the ratings. The same process is repeated for the physical examination checklist with a separate rating form for physical examination.

The same process is repeated for each station and its skill component in the test. This includes high and low video performances, orientation to the checklist and to the scoring. To maximize the outcome of the standard-setting procedures it is possible to divide, after the orientation, a large group of panelists (i.e. 18) into three groups of six each. The groups will set standards on different stations but one or two stations will be rated by all. This is done to examine the consistency of ratings among panelist groups.

### History scores

<table>
<thead>
<tr>
<th>Score</th>
<th>No.</th>
<th>Cum.No.</th>
<th>%</th>
<th>Cum.%</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>*</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0.0</td>
</tr>
<tr>
<td>*</td>
<td>7</td>
<td>1</td>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1</td>
<td>5</td>
<td>0.6</td>
</tr>
<tr>
<td>*</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>1.3</td>
</tr>
<tr>
<td>*****</td>
<td>10</td>
<td>7</td>
<td>14</td>
<td>4.5</td>
</tr>
<tr>
<td>*****</td>
<td>11</td>
<td>8</td>
<td>22</td>
<td>5.2</td>
</tr>
<tr>
<td>*********</td>
<td>12</td>
<td>16</td>
<td>38</td>
<td>10.4</td>
</tr>
<tr>
<td>*********</td>
<td>13</td>
<td>17</td>
<td>55</td>
<td>11.0</td>
</tr>
<tr>
<td>*********</td>
<td>14</td>
<td>25</td>
<td>80</td>
<td>16.2</td>
</tr>
<tr>
<td>*********</td>
<td>15</td>
<td>26</td>
<td>106</td>
<td>16.9</td>
</tr>
<tr>
<td>*********</td>
<td>16</td>
<td>22</td>
<td>119</td>
<td>8.4</td>
</tr>
<tr>
<td>*********</td>
<td>17</td>
<td>22</td>
<td>141</td>
<td>14.3</td>
</tr>
<tr>
<td>*********</td>
<td>18</td>
<td>18</td>
<td>151</td>
<td>6.5</td>
</tr>
</tbody>
</table>

* **
- 19 | 3 | 154 | 1.9 | 100.0
- 20 | 0 | 154 | 0.0 | 100.0

Notes: Total number = 154. Scores are rounded to the next whole number.

### Ratings of communication skills

As stated earlier, the communication skill score did not vary substantially from one station to the other. Therefore, setting a communication standard by employing a holistic approach may be appropriate for this particular situation. Panelists do not consider a standard for each station, but rather set a point of ‘how much is enough’ on the scoring scale for the total test.

**Orientation.** Panelists are presented with test materials relevant to the communications component of the test. This includes the theoretical framework, the criteria, the training materials for the assessors (in this case the standardized patients) and the scoring scale. Panelists will view videotapes of a range of performances (high, middle and low) and the application of the scoring scheme to the performance will be demonstrated. The borderline group characteristics are discussed and listed with relation to the communication skills. Panelists will then be asked to state what will be the point on the scale at which the borderline candidate will demonstrate minimal competence for passing a station with no reference to a particular station. Inherent in this standard-setting approach is the compensatory orientation. The stability or consistency of the communication skills component over stations (for example, Cronbach α reliability 0.85–0.90) justifies such an approach. In this case panelists are not concerned with compensating poor performance in one station with high performance on the other since performance is generally stable over stations.

The first ratings are presented and averaged across panelists. Actual performance data are presented and discussion revolves around the rate of failure relative to the established standard. Panelists are then asked to provide a second rating. The second rating will constitute the final standard.
for the communication skills component. Candidates’ communication scores for each station will be averaged to produce a total test communication score. Candidates who exceed the standard will pass the communication component of the test.

Evaluation

The evaluation component of the standard-setting procedures is important from a number of perspectives:

1. To support the validity of panellists’ inferences and actions by demonstrating a defensible process.
2. To demonstrate the reliability of the ratings by indicating a decrease in the variability of ratings. This means that more consensus was gained through discussions of ratings and feedback from the other panellists. During the process of discussions, the panellists identify their own stringent or lenient tendencies and may correct their ratings in the iterative process of multiple ratings.
3. To find out the effect of the normative information on the panellists’ ratings and ensure that the criterion-references approach did not shift to a norm-referenced approach owing to high influence of the normative data.
4. To find out the effectiveness of the orientation material and process for determining the extent to which the panellists changed their previous conceptions as a result of their orientation experience.
5. To determine the extent to which the panellists were confident in the process and the resulting standards. Panellists’ confidence in the standard is another indication of the validity of the process. A broad representative sample of panellists constitutes collective expert judgements for justifying a standard. This is another aspect that supports the defensibility of the standard-setting approach.
6. To obtain feedback from panellists as to how the process might be improved. This is an important aspect of evaluation since standard-setting approaches in performance assessment in the medical profession are still in an evolutionary stage. Any feedback on the improvement of the process may assist in the construction of improved approaches to setting standards. As Linn et al. (1982) state: “Standard setting methods depend on human judgements and there is no single perfect method of setting them”. Therefore any improvement in the process will maximize the benefits.

Evaluation materials should include data on the first and second ratings of the panellists for each of the test components rated, which should demonstrate increased consensus of raters. It should also include a questionnaire administered to panellists at the end of the standard-setting process. Test developers should design the questionnaire according to the test needs with consideration of the six points stated above. In high-stakes examinations the aspect of evaluation could not be overemphasized. However, even in other settings such as medical school or courses the defensibility of the process must be demonstrated through the evaluation process.

Concluding comments

- Much work is still needed to establish effective standard-setting procedures. Often, these procedures are challenged for their arbitrariness and for the unstable standards over occasions, over panellists and over methods (Jaeger et al., 1996). However, the growing need for establishing standards of candidates’ performance generates innovative and creative procedures for standard setting, mainly for performance assessment.
- The length of procedures (for the OSCE it may take two days of panellists’ work), should also be considered and ways to shorten the process are needed.
- The internal process of standard setting may incorporate a built-in mechanism to identify test components that ‘behave’ differently or may result in ‘odd’ standards. This could serve as an opportunity to identify test materials with ‘flaws’. Comparison with the other test components may assist in rectifying the problem (Clauser et al., 1996).
- Further consideration must be given to fully compensatory models (Clauser et al., 1996; Burrows, 1999) in which test items or components are averaged to produce a test standard. Is the averaging an adequate procedure? Is the total test standard a simple average of the test component standards, or should it be a weighted average, or any other form of aggregation?
- Obtained standards should be checked against other information available on the test taker to ensure the validity and the reasonableness of the standard.
- Effective methods of training panellists to recognize borderline characteristics are essential if the Angoff approach is widely used.
- The more standard-setting procedures are applied to a variety of tests, the more we will enhance the practice of high-quality testing, and the higher will be the confidence in the testing of professional competences.
- As Jaeger et al. (1996, p. 80) state: “the state of the art, of standard setting for performance assessment, is far from a state of grace. Much work remains to be done.”

Notes on contributor

MIRIAM FRIEDMAN BEN-DAVID, formerly Co-Director of the Clinical Skills Certification Program at the Educational Commission for Foreign Medical Graduates, is currently a visiting professor at the Centre for Medical Education at the University of Dundee. Her main area of research is the evaluation of physician competences, in undergraduate and postgraduate education.

Notes

[1] The practice case was drawn from the University of Dundee Medical School Clinical Skills Centre OSCE bank.

References


