
# Standard Setting

## Determining the appropriate pass-fail score

C. B. Hazlett with grateful acknowledgement of contributions from
E. Skakun (University of Alberta)
C. van der Vleuten (Maastricht University)

# Learning Outcomes

**After this session, you will**

- understand the rational for using, and the principles underlying, various standard setting methodologies
- determine if the Faculty's present standard setting method is sufficiently adequate

# Have You Ever Had This Experience?

So many students failed the test you had to adjust their marks

Student was passed based on scoring with the OSCE answer sheet, but you felt she had not understood the purpose of the procedure tested in the station, nor what she had found

Almost everyone in the class did exceptionally well on the test but you felt that they weren't that competent

You decided to pass a candidate but you worried that he was not adequately prepared to care for patients

# Contradictions

Have we ever created tests of different difficulty dealing with the same content or skills?

Are all classes which have multiple teachers/sections

receive the same level of instruction?

Are all subjects and skills which students need to acquire of equal difficulty?

Are all assessment formats (MCQ, Essay, OSCE, etc.) of equal difficulty?

**Then does it make sense to define adequate competency to be a fixed value (e.g., 50% is set as a pass mark)?**

# Whom Should We Pass & Fail?

## How Should We Decide?

## Will our Decisions be Defensible?

IDEAL

# This Session Addresses the Following Questions

**What** is standard setting?

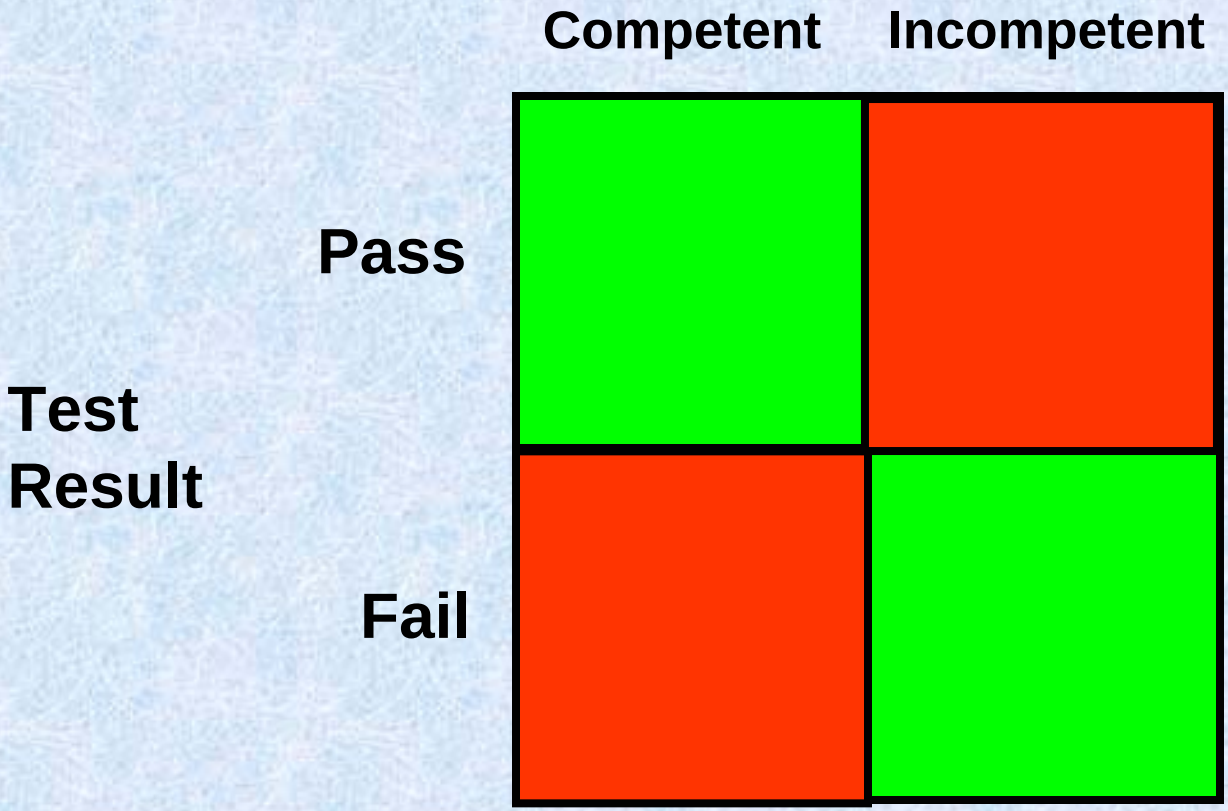
- **Why** is standard setting used?
- **Who** should pass and fail?

**Which** method of standard setting is preferable?

- **How** does one decide which method to use?
- **When** should standard setting be used?

**Where** can one find related evidence?

# The Standard Setting Problem



# Determining Pass/Fail Cut-off Points

**How much is enough to be adequately competent?**

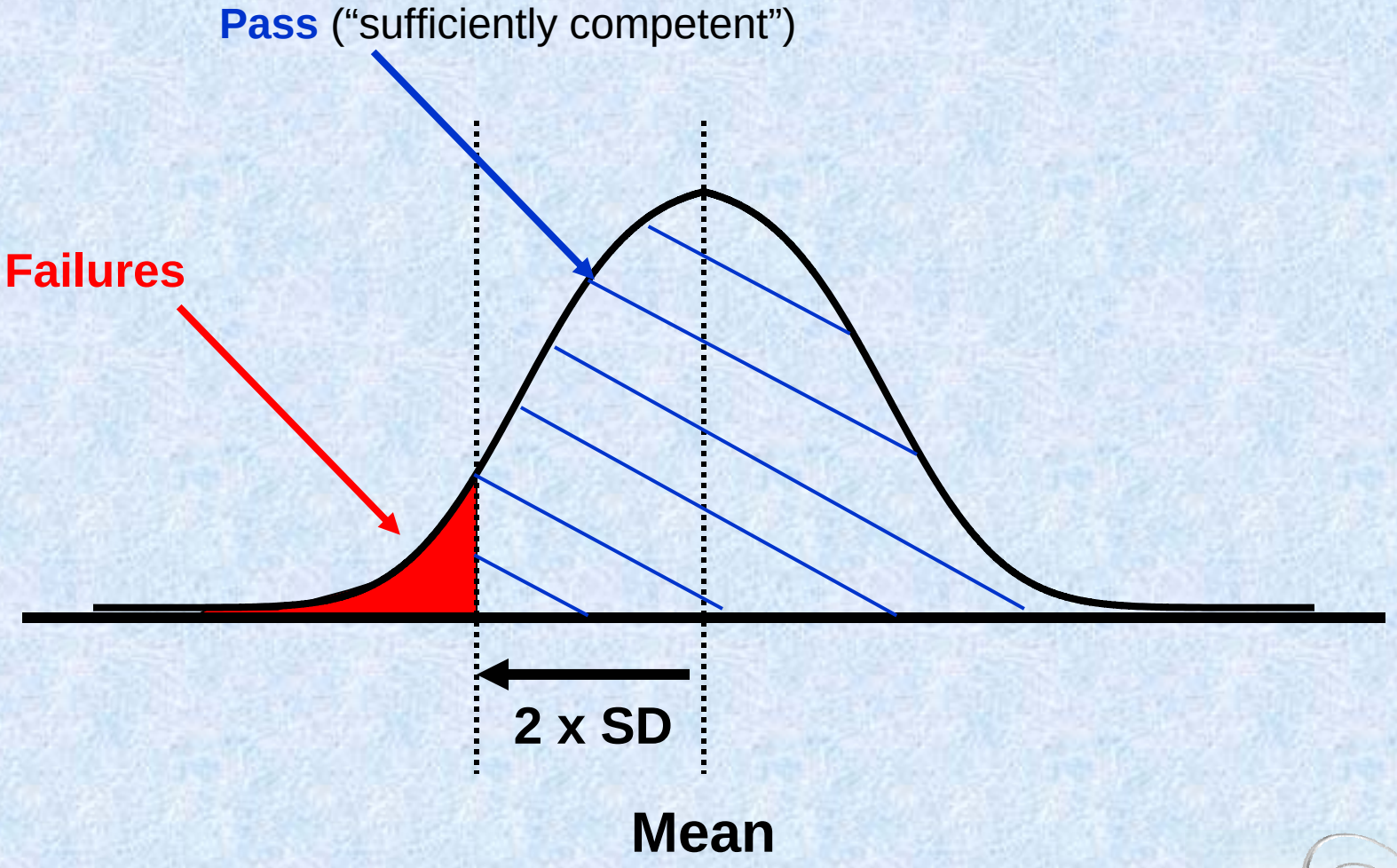
- the answer defines a passing score for an exam

Typical previous standard setting methods used to set the pass/fail point for a test have been

absolute & fixed: a particular score or a % which has been determined prior to the test is set as the pass mark: e.g. 50%

relative & not fixed: students are compared with each other & those who fail are “X” SDs below the mean performance of all candidates

# A Common Previously Employed Relative Method



# Problems with These Historical Approaches

## Absolute & Fixed Standard

- passing score is dependent on minimal mastery of study content
- but the minimum mastery point is determined *a priori* (e.g., by a school-wide policy), and thus
- ignores error variance due to unwanted variation in the quality of teaching & the test

## Relative & Non Fixed Standard

- passing score is dependent on performance of the reference group
- can correct for variation in teaching & assessment quality, but
- ignores error variance due to sampling (the reference group), e.g.,
  - some <u>below</u> $P_{2.5}$ might have scored over 80% on the test
  - some <u>above</u> $P_{2.5}$ may have scored even less than 35% on the test

# Pros & Cons of Absolute & Relative Standards

| | Absolute Standard | Relative Standard |
|---|---|---|
| Easy to defend | + | - |
| Corrects for lacks in quality of training and/or test | - | + |

# Present Practice

Given limitations of previous approaches, increasingly, the standard setting methods adopted by many medical schools & licensing bodies are

- test-centered or examinee-centered (or a combination of both)
- with the judgment of subject matter experts factored into the method of choice

There are now over 50 different such methods, only a few of which are discussed here

# Example Test Centered Standard Setting Methods

## Judgement of a test's items or OSCE's stations

a. Angoff *
b. Modified Angoff *
c. Nedelsky
d. Ebel

* Simply highlight as a review as already discussed in previous retreat

# Examinee Centered Standard Setting Methods

**Judgement of Individuals**

- Borderline
- Contrasting groups

**Judgement of Groups**

- Wijnen method
- Cohen method

**Compromise** (uses aspects from both test & examinee centered)

- Hofstee

# Numerous Standard Setting Methods

**Absolute Standard**

Judgement of Tests

- Angoff
- Nedelsky
- Ebel

**Relative Standard**

Judgment of Individuals

- Borderline
- Contrasting groups

Judgement of Groups

- Wijnen method
- Cohen method

**Compromise**

Judgment by Compromise

- Hofstee

Will highlight features of those noted in red

# Which Method Is Best?

**Gold standard is almost always unavailable**

- arbitrariness cannot be prevented or avoided

- credibility is therefore the key criterion for deciding which method is most appropriate for one's programme

IDEAL


# Example of a Test Centered Method

## in

## Standard Setting

# Angoff Method
# (Has been often used in Medical Education)

A group of experts pass judgment on the proportion of minimally competent (borderline) candidates who *could* correctly answer an item (or could correctly perform a procedure in an OSCE station).

- evidence: use a minimum of 8 judges (but 12-18 judges are needed to be safe: Margolis et al)
- each judge must keep in mind a commonly agreed upon definition for a hypothetical borderline candidate
- judges' estimates are averaged for each item
- cutoff point is set as sum of these averages.

# Angoff Illustrated

**A previously healthy 20 yr old female presents with sudden pleuritic pain in the left chest with SOB.**

**What is this lady's most likely condition?**

**A Mycoplasma pneumonia**

**B. Spontaneous pneumothorax**

**C. Pulmonary embolus**

**D. Acute pericarditis**

**E. Myocardial ischaemia**

Each judge specifies his/her estimate of the proportion of minimally competent candidates that can answer this item

# Estimates of Six Judges for the First Item in the Test

| Judge | Proportion of borderline candidates that should answer this item correctly |
|---|---|
| 1 | .85 |
| 2 | .80 |
| 3 | .80 |
| 4 | .95 |
| 5 | .85 |
| 6 | .90 |
| Average for item | .86 |

# Angoff Illustrated (cont'd)

The foregoing tabulation is repeated for each item in the test

i.e., the average of all judges for each item

The sum of these averages is the minimum pass mark

Assume the assessment had 12 items: see next slide

# Ratings by Six Judges for Entire Test of 12 Items

**Sum across 6 judges' estimates for each of the 12 items & divide by 6**

**Table 1** Example of the results of an application of Angoff's method by six judges to a 12-item test

| Question | Judge 1 | Judge 2 | Judge 3 | Judge 4 | Judge 5 | Judge 6 | Average |
|---|---|---|---|---|---|---|---|
| 1 | 0.85 | 0.80 | 0.80 | 0.95 | 0.85 | 0.90 | 0.86 |
| 2 | 0.60 | 0.70 | 0.50 | 0.55 | 0.65 | 0.70 | 0.62 |
| 3 | 0.45 | 0.55 | 0.50 | 0.60 | 0.90 | 0.35 | 0.56 |
| 4 | 0.90 | 0.95 | 0.90 | 0.95 | 0.85 | 0.90 | 0.91 |
| 5 | 0.80 | 0.75 | 0.65 | 0.70 | 0.85 | 0.55 | 0.72 |
| 6 | 0.70 | 0.65 | 0.60 | 0.70 | 0.75 | 0.60 | 0.67 |
| 7 | 0.40 | 0.50 | 0.35 | 0.50 | 0.55 | 0.50 | 0.47 |
| 8 | 0.75 | 0.65 | 0.60 | 0.70 | 0.75 | 0.60 | 0.68 |
| 9 | 0.65 | 0.55 | 0.70 | 0.65 | 0.65 | 0.60 | 0.63 |
| 10 | 0.55 | 0.50 | 0.45 | 0.60 | 0.65 | 0.55 | 0.55 |
| 11 | 0.50 | 0.45 | 0.40 | 0.50 | 0.55 | 0.50 | 0.48 |
| 12 | 0.95 | 0.95 | 0.95 | 0.90 | 0.95 | 0.95 | 0.94 |
| Cutpoint | | | | | | | 8.09 |

**Then sum the averages**

**Pass Mark is then at 8.09 out of 12 items or 67.4%**

Norcini J. Setting standards on educational tests. *Med Education*, 2003; 37:434-469.

# Angoff: Disadvantages & Concerns

Getting experts to agree & set standards is not easy

Can be time consuming for long tests

Judgment is based on hypothetical students, not on actual candidates in the examination

# Use Modified Angoff to Circumvent Some Concerns

**When there is disagreement among the independent ratings of the experts, these are discussed by the all the raters**

- individuals might subsequently decide to adjust some of their judgments; and/or

**Item Analyses of the student performances are considered**

- a further adjustment might be made by some judges

**If one or two judges remain as outliers to all the others**

- their respective ratings are dropped

# Examples of Examinee-centered Methods

in

## Standard Setting

# Judgment of Individuals Performances

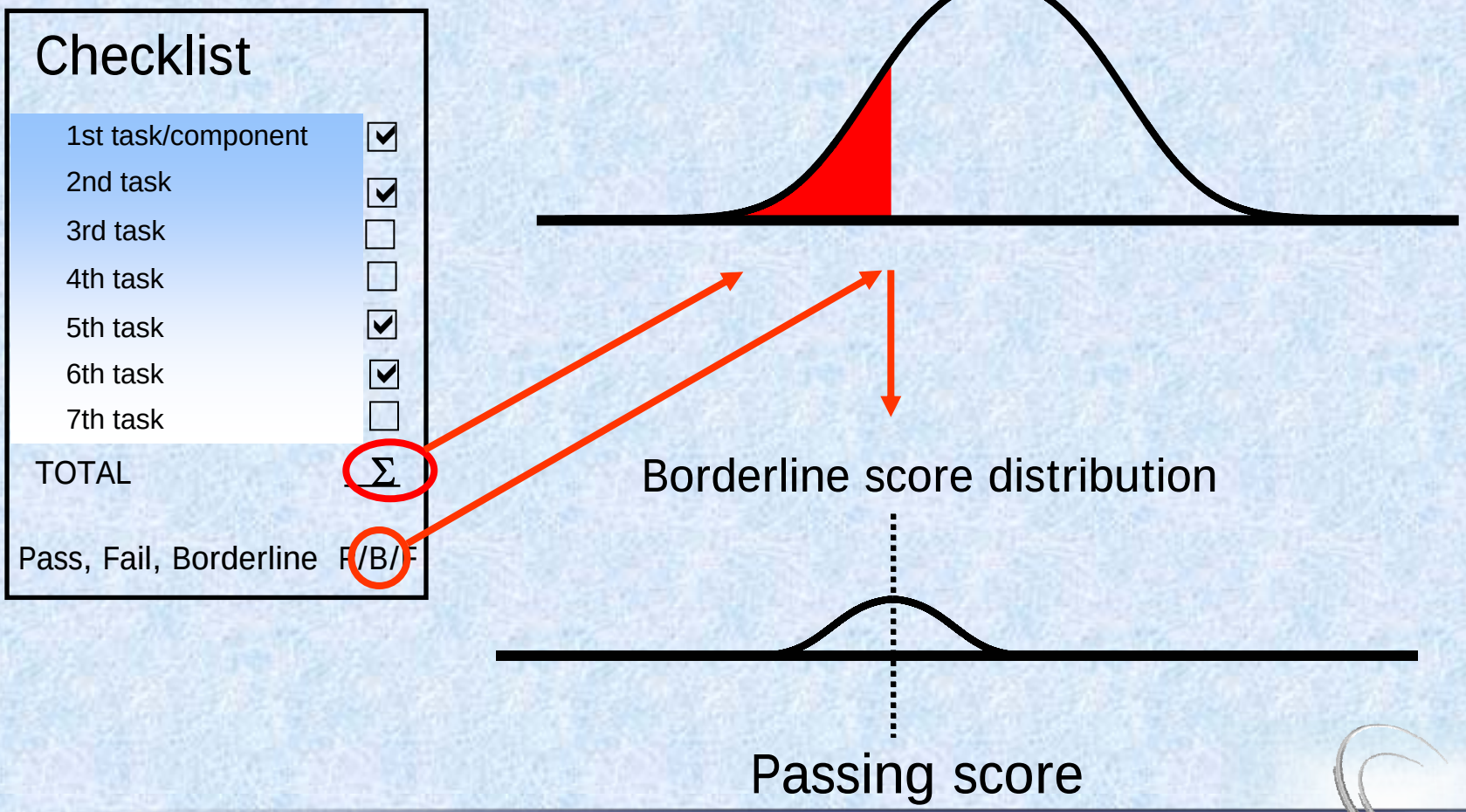
**Borderline group method**

**Contrasting group method**

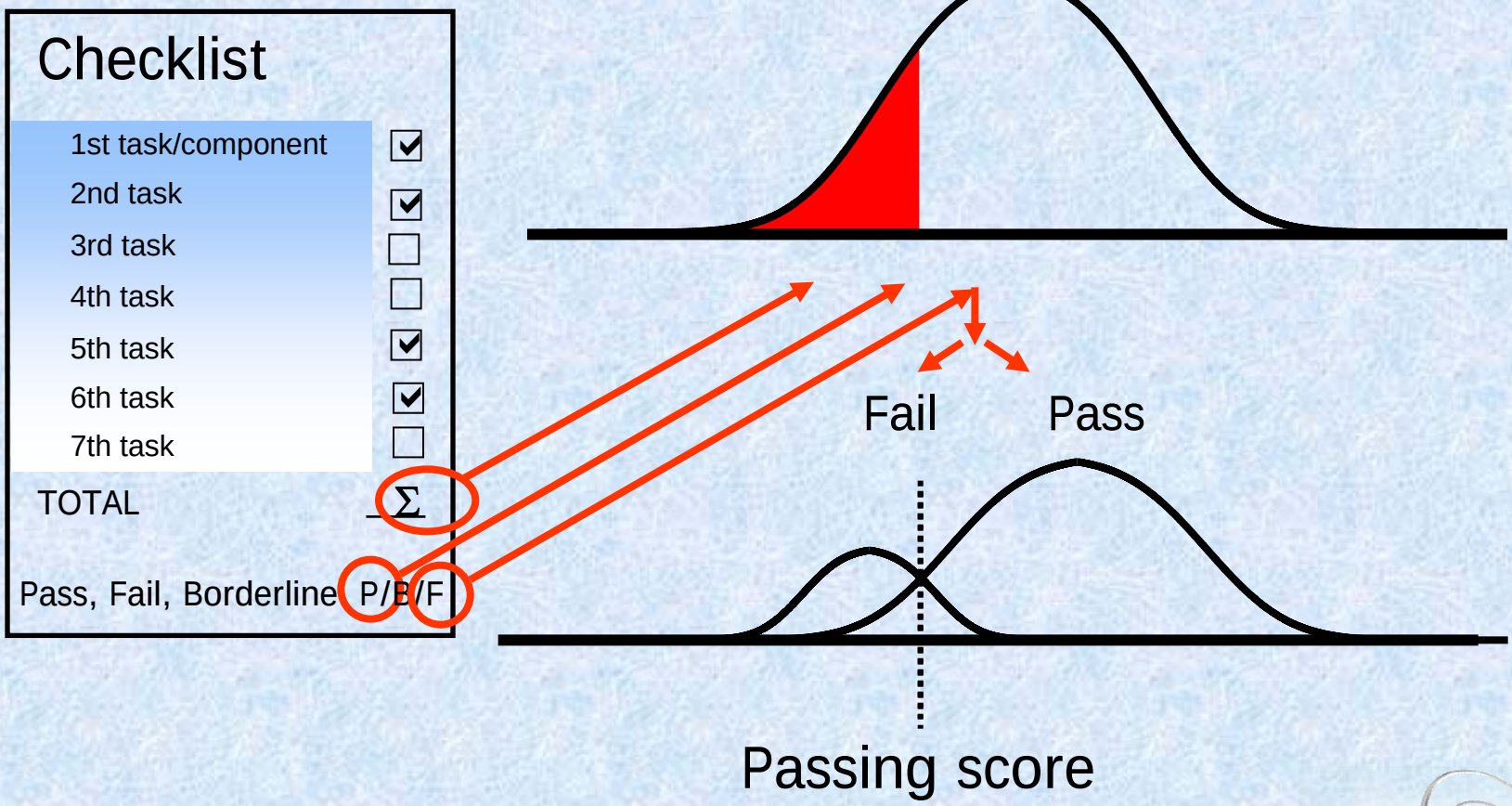
**Regression based method**

Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, van der Vleuten C
Comparison of a rational and an empirical standard setting procedure for an OSCE,
*Medical Education*, 2003 Vol 37 Issue 2, Page 132

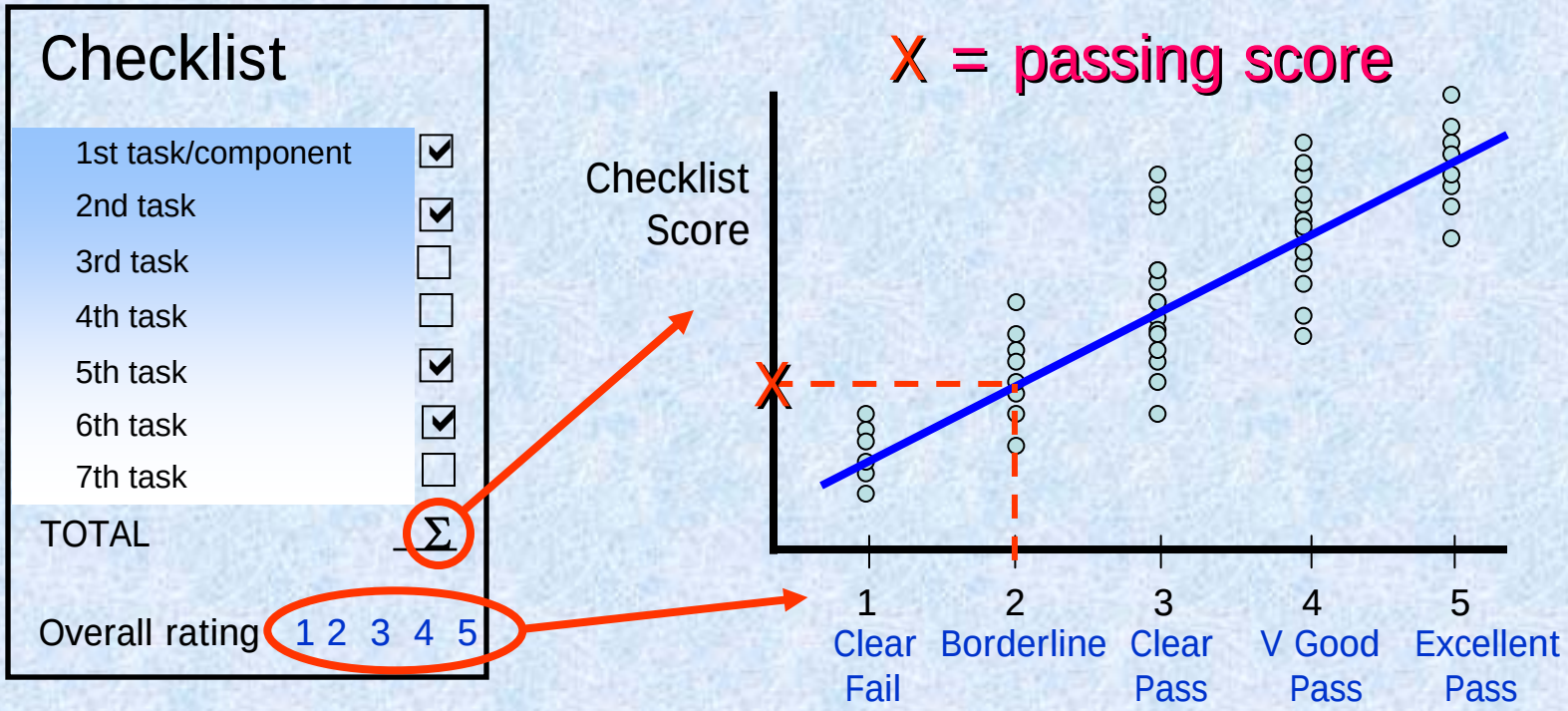
# Borderline Method for Standard Setting



# Contrasting Groups Method for Standard Setting



# Regression Based Method for Standard Setting

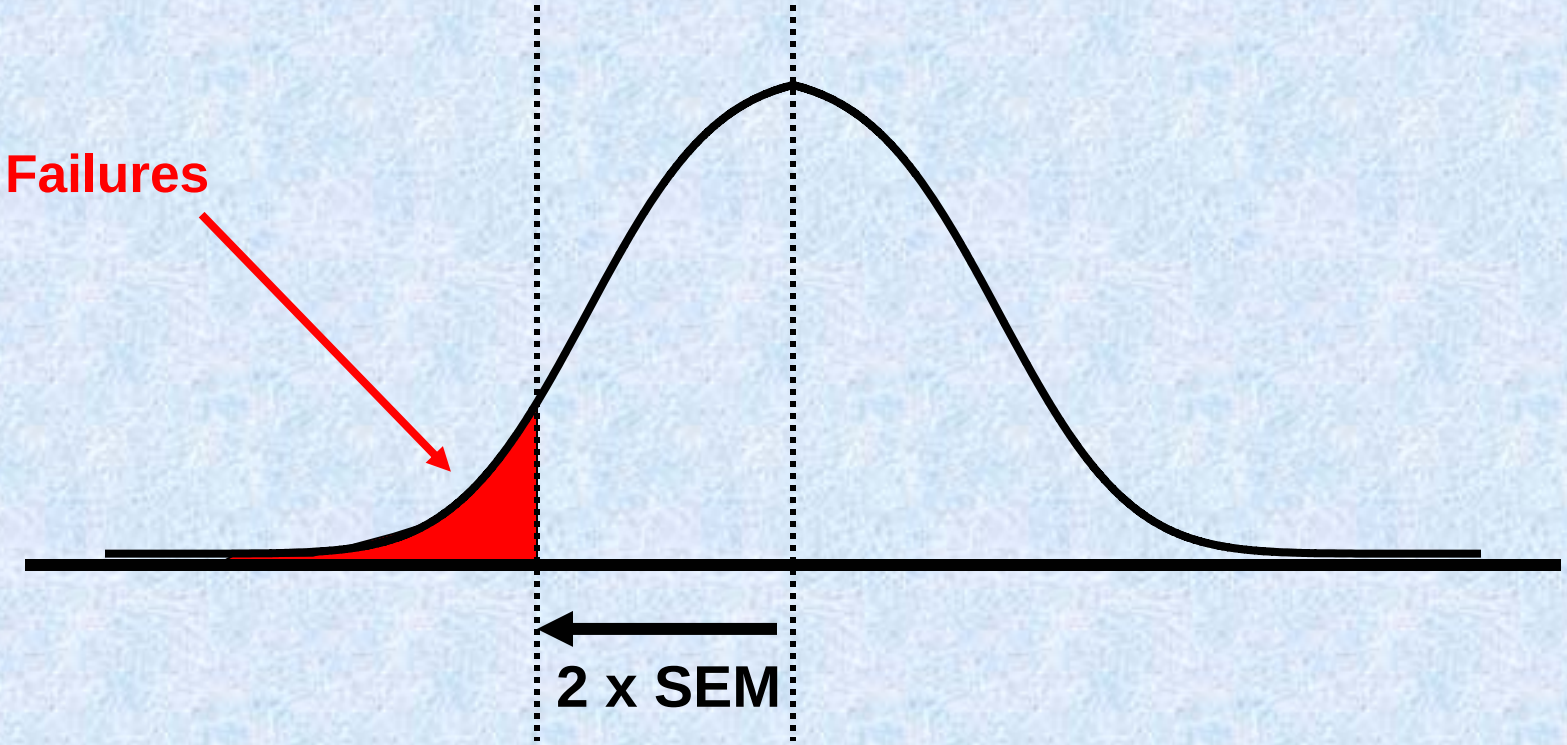



1=Clear Fail 2=Borderline 3=Clear Pass 4=Very Good Pass 5=Excellent Pass

# Borderline / Contrasting / Regression-based Methods

- **Not difficult to obtain a panel of judges (judges are the examiners)**
- **Reliable: cut-off score based on large sample of judgments**
  (no. of stations x no. of candidates)
- **Credible: based on expert judgment in direct observation**
- **Passing score not known in advance**
  (as occurs with all examinee centered methods)
- **Judgments not independent of checklist scoring**

# Wijnen Method: Judgment of Groups



# Advantage of Wijnen Method vs Typical Relative Method

**Using the Standard of the Mean (SEM), it is recognized that**

the typical performance estimate (mean) will vary due to sampling error

the lower bound of the 95% Confidence Interval for the mean could be “typical”, but below that the performance is unlikely “typical”, i.e., is probably “not assuredly adequate enough”

# Practical Implications

**Choice of standard setting methods depends on:**

- Credibility
- Resources available
- Importance of the high stake test

# Conclusions

Be aware of substantial “noise” in decision making

Substantial variation in instructional and test quality exists (for the latter particularly because of the difficulty of writing high quality items)

The best standard does NOT exist; every standard is arbitrary

A good standard is a credible standard

A standard which takes variation of educational quality into account is more credible than a standard which doesn’t

# Conclusion

Regardless of the method used to set a pass-fail standard, the judgment will be necessarily arbitrary

But the judgment should NOT be capricious

- use of expert judgment by a group of wise women & men in conjunction with
  - their discussions to reach a consensus and
  - the analysis of student performances on the exam
- is not viewed as capricious by accreditation bodies, students, legal experts, or the public for whom students will eventually be providing medical care

# The Paradox & The Reality

## Setting standards absolutely requires some relativity

Cees van der Vleuten

IDEAL

# Suggested Reading & Resources


Cusimano MD. Standard setting in medical education. *Acad Med* 1996;71(10 Suppl):S112-20.

Livingston SA, Zieky MJ. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton NJ: Educational Testing Service, 1982.

Norcini JJ, Shea JA. The Credibility and Comparability of Standards. *Applied Measurement in Education* 1997;10(1):39 - 59.

Cizek GJ (Ed.) (2001) Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum.

van der Vleuten C Overview on several standard setting methods available at http://www.fdg.unimaas.nl/educ/cees/oman

IDEAL


# Q & A & Discussion

Should we adopt alternate standard setting methods for our student assessments?
If so, how should we decide which method to use?
Should standard setting be used for all assessments or only some (e.g., end of module/panel/year assessments)?