Using Cross-Classified Models to Determine Between-Rater Variance in Standard Setting Ratings

Drs. Bridget McHugh, Dave Glerum, & Rebecca Berenbon

Today's Agenda

- Background on the Org & Project
- Overview on Standard Setting
- Rater-Specific Variance and Standard Setting
- Overview of Cross-Classified Random Effects Modeling
- How to Conduct CCREM in SAS, R, and SPSS
- Demonstration of Method
- Possible Causes and Solutions for Rater-Specific Variance
- Questions

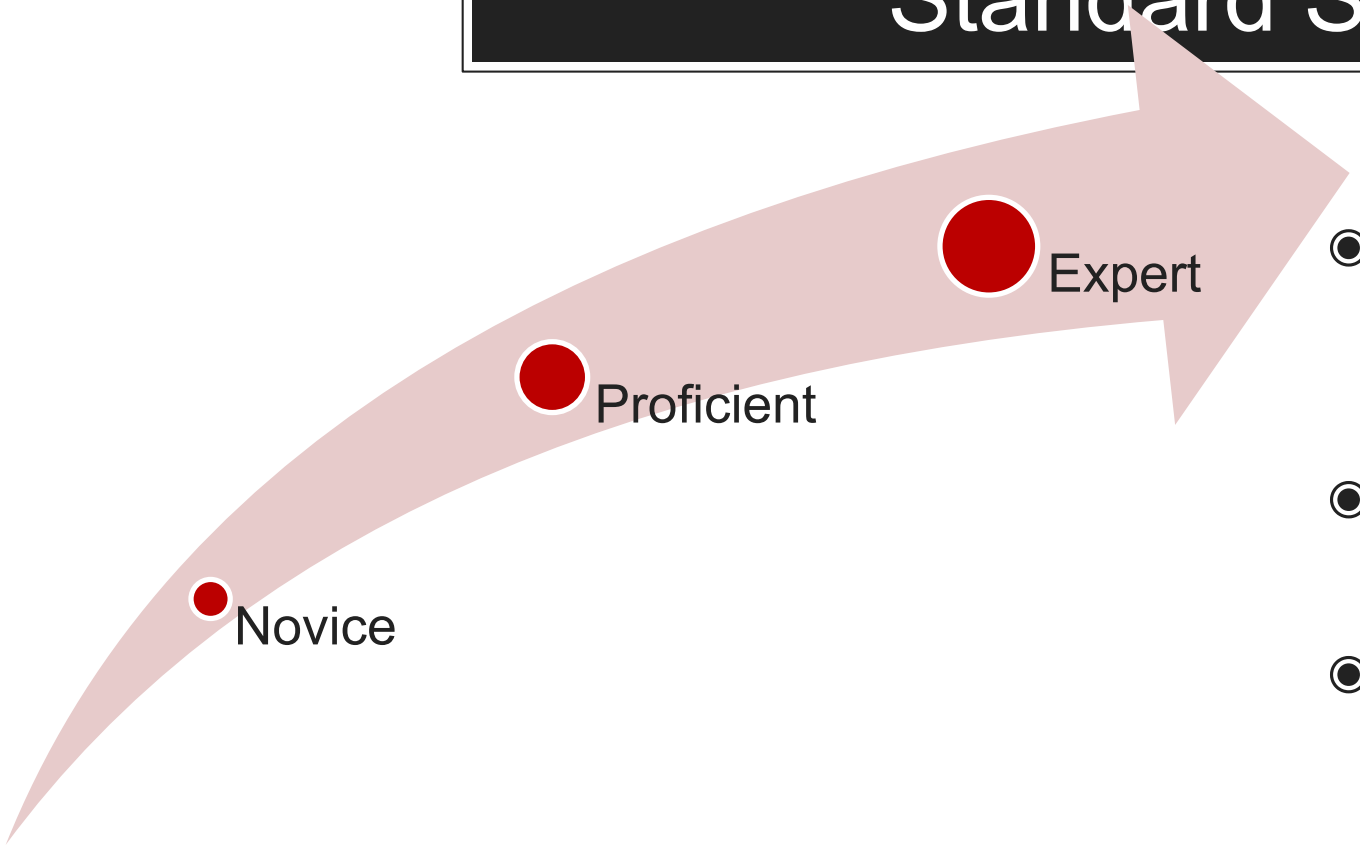
What is CETE?

- Focuses on translating research into practices that innovatively address critical issues
- Assessment, program evaluation, training, curriculum development
- Mix of I/O Psychology, Educational Measurement, and Instructional Design backgrounds

Standard Setting

the methodology used to
define levels of achievement or proficiency and
the cut scores corresponding to those levels

Standard Setting



- Essentially, tells us what score is used to establish acceptable or **minimal proficiency**
- Could be multiple (e.g., novice to expert)
- Used in **criterion-referenced** rather than norm-referenced test in:
  - Hiring
  - Credentialing (certifications, etc.)
  - Some educational contexts

Standard Setting

- Angoff
  - Proportion that would get correct
  - Yes/ No (what we use)
  - Essentially based on likelihood of getting question correct
  - May be done in several rounds
  - Time-intensive
- Bookmarking
  - Order based on difficulty, pick the “borderline” question
  - Requires data
- Direct consensus
  - Based on judgement of the overall test
  - Less intensive, but what if you add/drop items?

Standard Setting

Why it's important to have an accurate performance standard

- Too high– label proficient people as non-proficient
  - Indicate a skill gap when there is none
  - Consequences for program if done for accountability purposes
  - Artificially restricts your applicant pool
- Too low – label non-proficient people as proficient
  - Consequences for community for credentialing (e.g., licensing)
- Both are bad because they reduce trust in the assessment

Rater-Specific Variance

- Problem– different people have **different ideas of "minimum proficiency"**
- **Rater errors** can happen in standard setting too
  - Leniency
  - Severity
  - Central tendency
- Errors can also result in restriction of range **across tests**
- Can see how much variance comes from the individual rater's personal schemas and rating errors / idiosyncrasies by looking at ratings across several tests

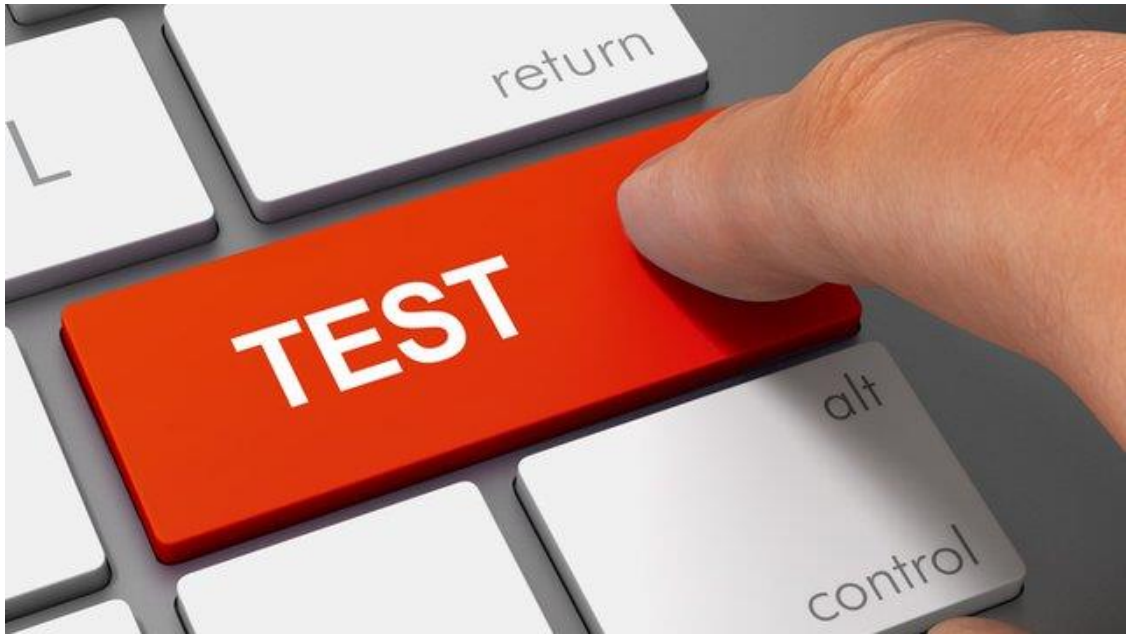
Cross-Classified Modeling

Form of multi-level modeling that accounts for forms of nesting that are not within one another (i.e., not just hierarchical)

Cross-Classified Models

- An underutilized but helpful form of **multi-level modeling**
- Appropriate when nesting structure doesn't neatly "fit" into each other in a purely hierarchical fashion
- Example of purely hierarchical nested structure: team → immediate supervisor → unit lead → department head
- Example of nested structure appropriate for CCREM: cross-functional team, each member has different work functions (e.g., sales, IT, administrative), some have additional projects based on individual skills regardless of function each with separate project lead
- In above example, these categories are **not mutually exclusive**

Demonstration Assessment



- Career skills tests
- Each test has different levels of skill (e.g., 100, 200, 300 level)
- SMEs are instructors in each course
- Courses are nested within career track, each instructor teaches different classes

Demonstration Data Set

- *Min*= 4 and *Max*= 19 instructors, *Mean* = 9 per a course, across 43 raters
- 24 tests included in standard setting process
- Each rater provided standard setting ratings for an average of 6 courses
- Modified versions of the Hofstee compromise and Beuk adjustment
  - What do you think passing score should be?
  - Min and max expected to pass?
  - Min and max acceptable passing score?
  - Min and max you could tolerate?
- RatingQues was turned into ordinal variable, with lowest = minimum tolerable score, highest = maximum tolerable score (confirmed by mean values)
- Level = course progression in series

CCREM Application

- Need to know **where variance in ratings stems from**
- CCREM can disentangle **irrelevant sources of variance** (i.e., rater-specific variance) from **relevant sources of variance** (i.e., pathway of course, level of course)
- Can base cut score on the method with lowest rater-specific variance
- Not: generally use no more than 3 effects

How to Do in SPSS

- Use **mixed models**
- DV is rating, IV is **course**
- Separate model for each rating
- Nesting variables are:
  - Rater
  - Class Level (100, 200, 300)
  - Class Pathway
- **Most variance should not come from rater**
- You can use VARCOMP with interactions but takes a long time

mixed rating with course
/fixed course ratingques
/print = solution
/random intercept | subject(rater)
/random intercept | subject(level)
/random intercept | subject(path).
EXECUTE.

How to do in R

- Use lme4 package
- DV is rating, IV is **course**
- Separate model for each rating
- Nesting variables are:
  - Rater
  - Class Level (100, 200, 300)
  - Class Pathway
- **Most variance should not come from rater**

```
cutscores = lmer (
        rating ~
        course ratingques
        + (1|rater)
        + (1|level)
        + (1|path)
)
summary(cutscores)
#last line gives variance components
```

How to do in SAS

- Use mixed procedure
- DV is rating, IV is **course and**
  Separate model for each rating
- Nesting variables are:
  - Rater
  - Class Level (100, 200, 300)
  - Class Pathway
- **Most variance should not come from rater**

```
proc mixed data =ssdata covtest noclprint method=ml;
   class rater level path;
   model rating = course ratingques / solution ddfm =satterth;
   random intercept / subject=rater;
   random intercept / subject=level;
   random intercept / subject=path;
run;
```

Results

- Rank each rating on rater-specific variance
- Other sources of variance:
  - Course level
  - Path
- Do follow-up analyses, such as models with specific rating question to determine question with lowest variance due to rater
- Method with lowest rater variance is best (range is118-305 in example)
- Can also do interaction effects

Parameter	Estimate	Std. Error
Residual (e)	95.894	3.264
Rater (u_{01})	147.618	32.81
Level (u_{02})	1.295	0.82

- As you can seen above, rater has a large effect on ratings, especially compared to other grouping variables (level)

What is minimally proficient?

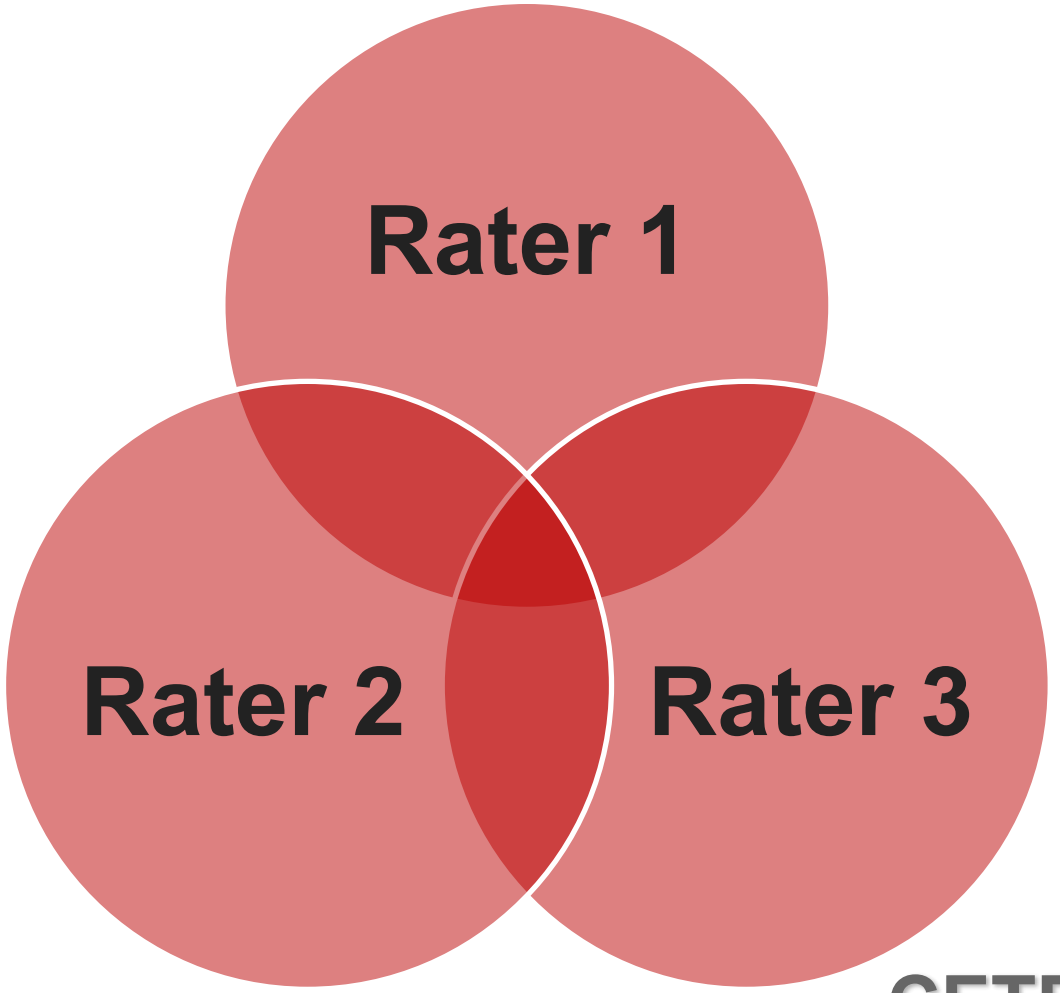
- **Personal schemas** for what is considered “minimally proficient”
  - Ability distribution of test takers
  - Difficulty of the test
  - Impact of standards on population
  - Consequences of passing scores
- Highly influence by **personal experiences and knowledge**
  - if minimal competence cannot be clearly defined, other factors are more likely to influence ratings
  - interpretations of learning objectives
  - more knowledgeable judges recommend higher passing scores

What is minimally proficient?

- Personal attitudes about assessment including:
  - Tougher graders
  - Like to assess people in relation to their peers
  - Interpret items in terms of general skills, may see items as easier

Interrater Consistency

- High consistency
  - the same level of difficulty receive similar ratings from a single judge
- Some previous work suggests consistency influence by standard-setting technique (Chang, 1999)
- Easier items have more consistent item-level ratings



Solutions

- Discussion between rounds of ratings with discussion to build consensus
- Use wording that more easily relates to past experience (e.g., the C student rather than the A/B student)
- Rater training to establish a shared mental model of minimal proficiency

Solutions

- Switching from test-centered to examinee-centered methods
- Carefully screening and selecting raters for inclusion in the standard setting study
- Conducting outlier analyses to determine if the between-rater variance for the same course is due to one extreme rating from a single rater
- Asking SMEs to take the test themselves (though may not work for very knowledgeable SMEs)
- Providing test-taker data (if possible) so raters base ratings on actual item difficulty, not where they would get it correct

Possible Applications

TRACKING TRENDS

Once we account for interrater variance, can see how time impacts rating scores across different forms of the test

TARGETED TRAINING

If we know which cut score methods have low agreement, can tailor training and discussion between rounds to the issue.

INDIVIDUAL SME COACHING

After determining variance, follow-up analyses can be used to determine which errors each SME is committing

INDIVIDUAL DIFFERENCES

Look at individual differences driving rater variance

CHECK CONSENSUS

Individual rater variance should shrink between ratings rounds if they are helpful / build consensus

SELECTION OF CUT SCORE

Pick cut score method based on which one has the least interrater variance through interaction effect

Questions?