Influences of IRT Item Attributes on Angoff Rater Judgments

Christian Jones, M.A.
CPS Human Resource Services

Greg Hurtz, Ph.D.
CSUS, Sacramento

CPS
Human Resource Services

Angoff Method

- Assemble a panel of subject matter experts (SMEs)
- Instructed to make item level judgments based on the performance of a consensually defined minimally qualified candidate or MQC (Angoff, 1971; Livingston & Zeiky, 1982).

Minimally Qualified Candidate

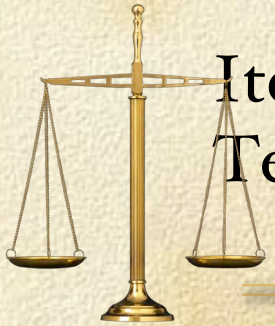
- Concept describes what the minimally competent person should be able to do, or know, the first day on the job
- Defined according to observable work behaviors that demonstrate his/her KSAs.
- Judgments are largely based on the comprehension of an MQC.
- Training is necessary to dissuade individual cognitive processes that may produce individual differences and variability of an MQC definition (Fehrmann, Woehr, & Arthur, 1991; Maurer & Alexander, 1992).

Angoff Method (cont.)

- Upon consensual MQC definition, SMEs consider each item and judge the probability that a borderline test-taker (MQC) would answer a test item correctly.
- Estimates are aggregated in the form of percentages or probabilities (p-values) across items for each judge to determine cutoff score.
- Method is appealing largely because the apparent ease of making simple probability estimates, and subjectivity is not obscured.

Item Statistics and Item Judgments In Classical Test Theory and Item Response Theory

- **Classical Test Theory (CTT)**
  - **Item Difficulty**- Proportion of MQCs who score correctly on an item.
    - Assigned an item-difficulty index or p-value ranging from .00 - 1.00; optimal values range from .30 - .70.
  - **Item Discriminability**- a measure of the effectiveness of an item at discriminating between high and low ability candidates.
  - **Guessing**- the probability that a candidate will answer a question by chance alone.



Item Statistics and Item Judgments In Classical Test Theory and Item Response Theory (cont.)

- **Item Response Theory (IRT) 3-parameter model:**
  - **Item Difficulty index (*b-parameter*)**- point on the ability scale at which the probability of correct responses to the item is .50
    - Theoretical range: $-3 \leq b \leq +3$ (Baker, 2001)
    - The *b-parameter* identifies the ability or theta (θ) value at the point where the slope is the highest or steepest for the item.

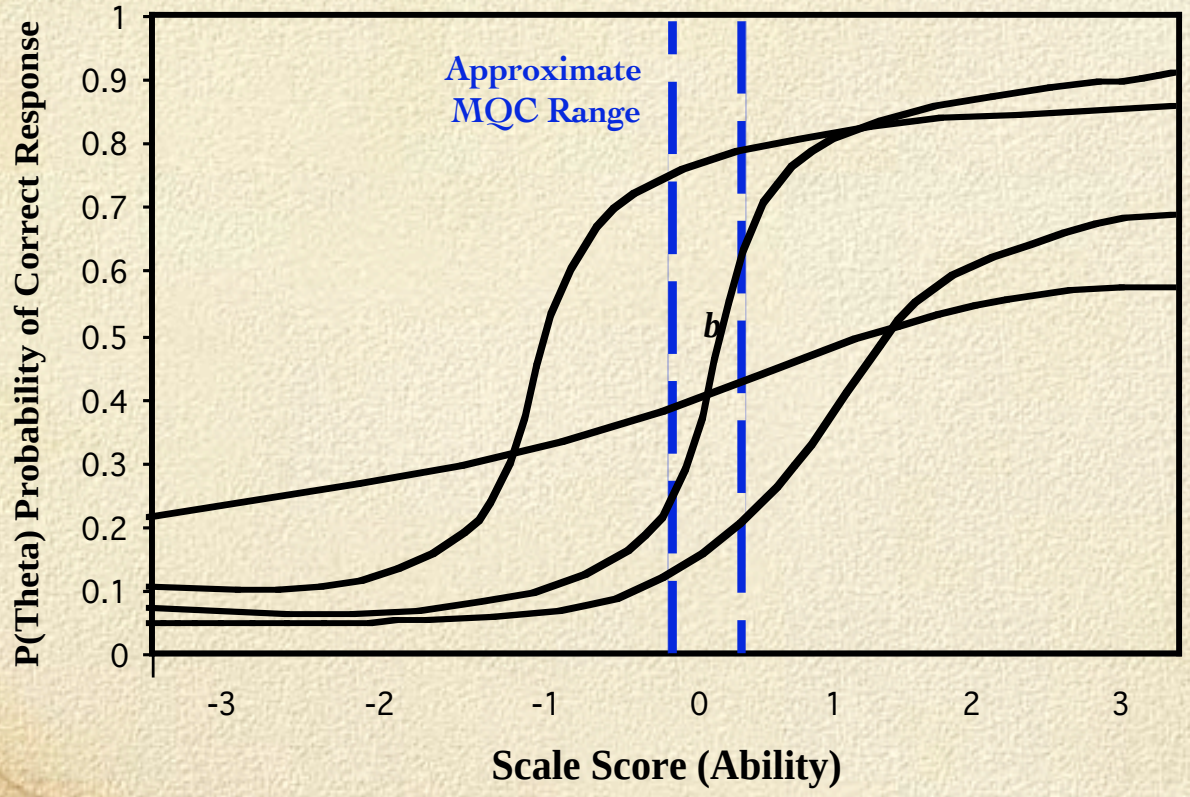
Item Statistics and Item Judgments In Classical Test Theory and Item Response Theory (cont.)

- **Item Discrimination index (*a-parameter*)**- describes how well an item can differentiate between examinees having abilities below the item location on the θ continuum and those above.
  - This parameter is proportional to the slope of the Item Characteristic Curve (ICC).
  - Theoretical range: 0 - 2.0
- **Pseudo-chance level index (*c-parameter*)**- candidate performance at the low end of the θ continuum or the probability of answering the item correctly by guessing alone.
  - Theoretical range: 0 - 1.0, but in practice, values above .35 are not considered acceptable.

Item Characteristic Curves

- An ICC provides a detailed map of item functioning across the entire range of θ or proficiency level.
- ICCs specify a relationship between the observable examinee item performance (correct and incorrect responses) and the unobservable traits or abilities (denoted by θ) assumed to underlie performance on the test.
- The task of an Angoff rater is related to each ICC.

Approximate MQC range, theoretical MQC range, and items functioning differently across the entire Ability continuum due to item discrimination, difficulty, and the pseudo-chance level.



Item Characteristic Influences* on Angoff Raters

* Assuming raters are sensitive to item properties when making judgments.

- ***b-parameter***
  - Raters must assess the item difficulty for candidates at a particular ability level (i.e., MQC definition).
- ***a-parameter***
  - May influence Angoff raters in their item judgments because a rater judges where (or how much) an item discriminates between the successful performing candidate and the unsuccessful candidate.
- ***c-parameter***
  - May influence ratings if raters accurately perceive the likelihood of candidates guessing the correct answer to a question.
  - **Advantages of using IRT for scoring and evaluating Angoff ratings hinge upon satisfactory fit between the model and the test data.**

Item and Rater Fit

- **Item Fit to IRT parameters**
  - Non-fitting parameters due to:
    - Wrong ICC model has been employed (1, 2, or 3 parameter models).
    - Values of observed proportions of correct responses are so widely scattered that a good fit cannot be obtained.
  - Poor fit results cannot yield invariant item and ability parameter estimates or ICCs- quality performance analysis is lost (Hambleton, Swaminathan, & Rogers, 1991).

Item and Rater Fit

Rater Fit to the IRT model

- Rater fit analyses should be performed to ascertain the most valid cutoff score and locate sources of variability in passing score estimates (Kane, 1987).

- Kane's chi-square analysis

  - Fit occurs when the MQC probability estimates for each item for the average rater approximate the value of $P(\theta^*)$, which is the height or steepness of the ICC at the MQC level on the θ scale, denoted, θ^*.

  - Translates the performance of an MQC into a cutoff score on the true score scale by utilizing the test characteristic curve (TCC).

Item and Rater Fit

- Van der Linden (1982) provided means to test whether a judge has rendered consistent, compatible judgments congruent with an MQC anchor (i.e., whether the rater has specified correct probabilities of success).

- **Intrajudge Consistency Analysis**

  - **Errors of Specification** analysis tests this assumption of consistency

  - **Index of Consistency** compares results between different judges or tests.

Item Judgments and Cutoff Scores

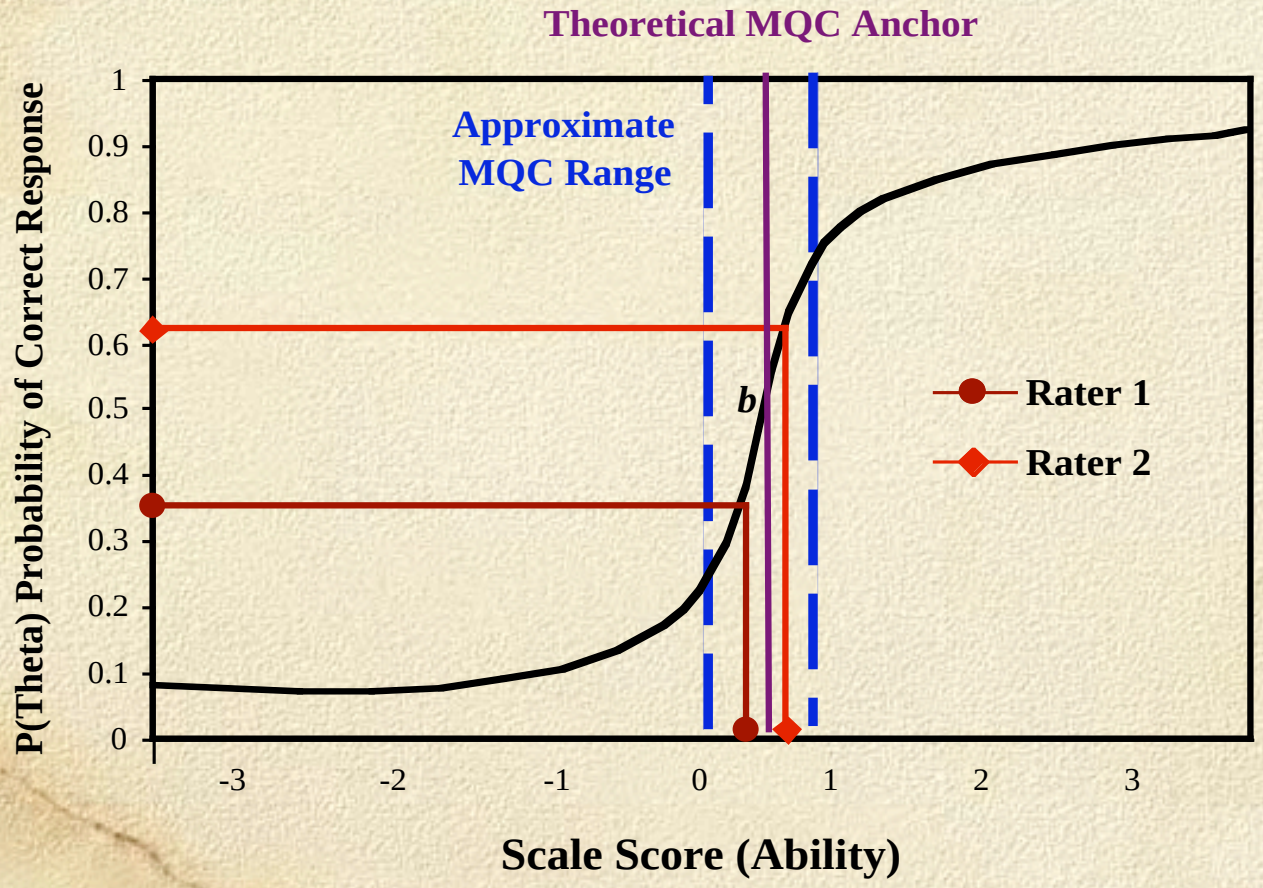
- CTT cutoff score using the Angoff method
- Cutoff score on the θ scale in IRT
  - The sum of the ICCs for each item in a given exam = Test Characteristic Curve
  - Assuming item parameters fit the data and raters fit the model.
- Kane's Method 2 (1987) utilizes the TCC in estimating the θ*.
  - Most commonly used procedure for deriving a passing score on a test from the MQC levels for individual items.
  - Precision of the θ* cutoff score translated through the TCC depends on adequacy of item and rater fit.

Hypotheses

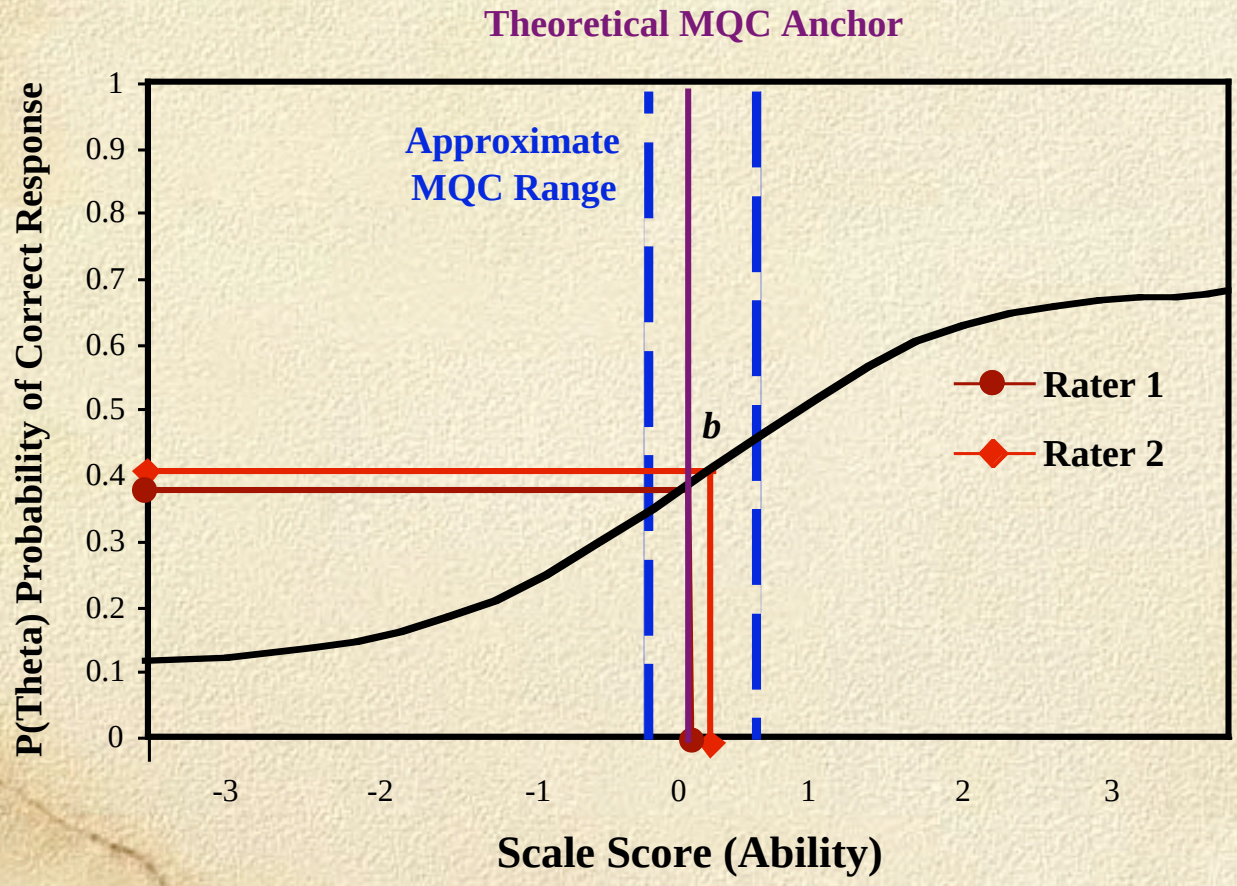
- **H1**: The slope (a-parameter or discrimination index) in an IRT item characteristic curve is correlated with the variability among Angoff ratings. It is proposed that higher discriminating items are more difficult for judges to agree upon; that is, raters will disagree more and vary in their judgments for items that have a steeper slope.

- Correlational analysis was utilized to determine the relationship between the slope and the variability of Angoff ratings.

Variance in Angoff ratings is influenced by individual perceptions of MQC definitions, the item's discriminating parameter, and location of the item's *b-parameter* near an MQC anchor .



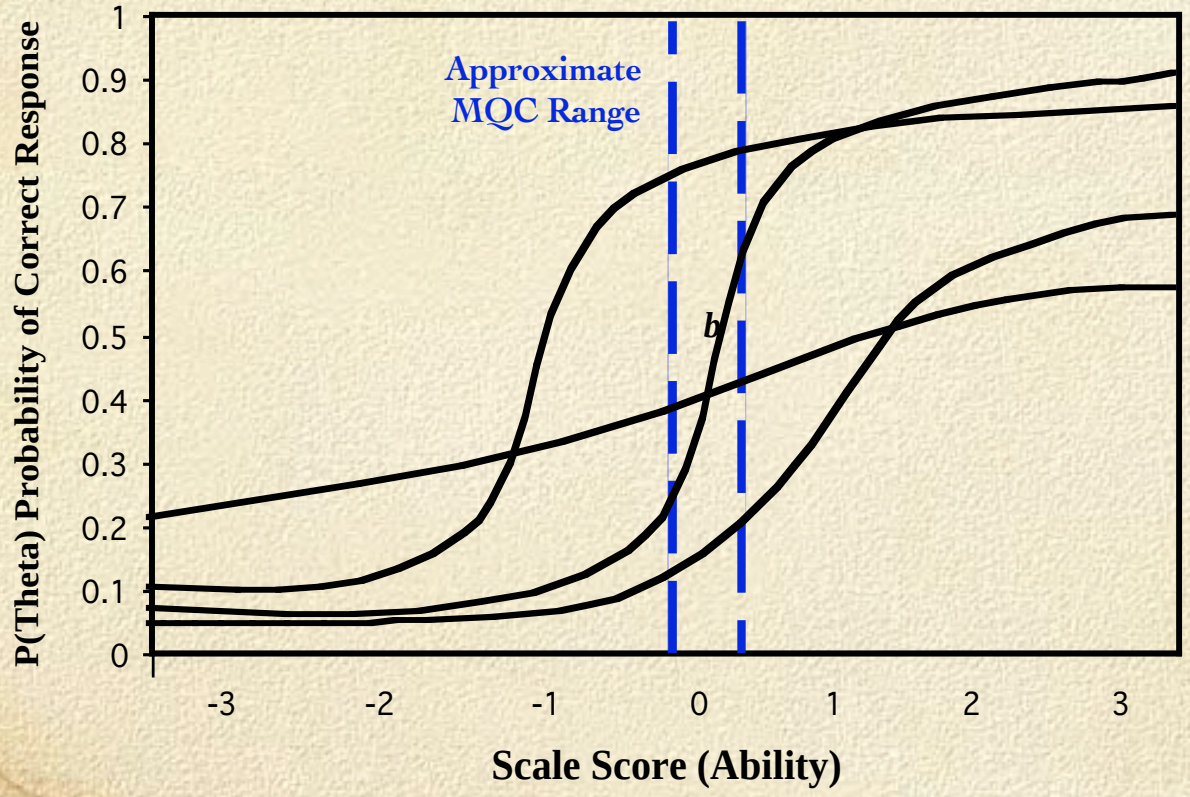
Similar rater item estimates vary only slightly due to the influences of the flatness of the slope (*a-parameter*), difficulty level, and MQC anchor.

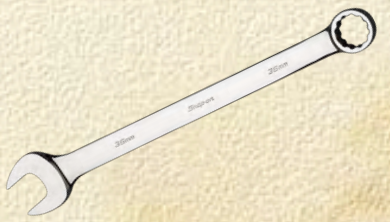


Hypotheses (cont.)

- H2: Assuming that a steeper slope yields increased variability of ratings in the approximate MQC range, it is hypothesized that this variability in ratings will occur only when the item's *b-parameter* falls near the MQC anchor or the data falls in line with the raters' MQC defintion.
- Multiple regression analysis will be utilized to determine this interaction.

Approximate MQC range, theoretical MQC range, and items functioning differently across the entire Ability continuum due to item discrimination, difficulty, and the pseudo-chance level.





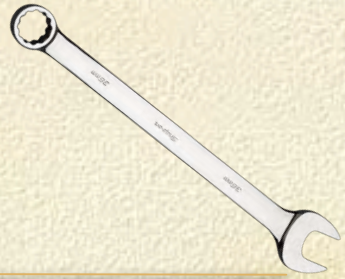
Methods

- Participants
  - Twenty-six licensed notary SMEs
- Materials
  - 174, four-option multiple-choice items taken from a State of California notary licensure exam.
  - Data represented statistical information collected over a three-year period of notary test development and administration from a total of 96,231 candidates.

Methods (cont.)

- Item characteristics were calibrated according to the 3-parameter model and ICCs were assumed to appropriately fit the model.
- The 3-parameter model was selected due to its its flexibility and ability to accommodate difficulty, discriminability, and the pseudo-chance level indices.

Methods (cont.)



- Procedure
  - All SMEs received frame-of-reference training from trained professionals.
  - MQC definitions to anchor raters
  - Modified Angoff method training and practice items.

Results

- After assuming appropriate item fit to the 3-parameter model, conformity of rater estimates to the ICCs was assessed.
- Out-of-range items were filtered out after SME judgments
  - *a-parameter* $0 \leq a \leq 2$
  - *b-parameter* $-3 \leq b \leq 3$
  - *c-parameter* $0 \leq .35$
  - Item sample size decreased from 174 to 130; ensured greater measurement accuracy and rater fit to the model.

Results (cont.)

***Advantages of using IRT for scoring and evaluating Angoff ratings hinge upon satisfactory fit between the model and the test data.**

- **Assessing Rater Fit**
  - Kane's Chi-squared analysis
  - Intrajudge consistency analyses

Results (cont.)

- **Assessing Rater Fit**
  - **Kane's Chi-squared analysis** (Kane, 1987; Hurtz, Jones, & Jones, manuscript in preparation) was employed to assess the relative goodness-of-fit of the raters to the data model for both the original 174 items and the 130 conditional item sets.
  - Both models showed poor indices of fit; therefore rejection of the null hypothesis was substantiated.
  - Improvement in rater fit is directly observable when out-of-range items were removed.

Results (cont.)

Intrajudge Consistency Analyses (Van der Linden, 1982)

- The **mean error of prediction** in the data set was .16, indicating that on average the judges' estimates differed from the ICC-based probability values by .16.
- The **mean consistency index** was .77 (the closer to zero, the less acceptable the hypothesis of a consistent judge.

Results (cont.)

Correlation

- **H1:** Slope in an ICC would be correlated with variability among Angoff raters.
- Results indicated a non-significant relationship between higher discriminating items and variance in item judgments.

Results (cont.)

Multiple Regression

- **H2:** Assumed that the relationship between the slope and variability in judgments would be more pronounced for those items where the point of maximum slope falls near the MQC.
  - Analysis revealed non-significant interactions.

Results (cont.)

- Non-rater fit to the ICCs and non-significance of expected relationships prompted further investigation into how individual Angoff judges might be influenced by IRT item attributes (*a-parameter, b-parameter, c-parameter*).

Results (cont.)

- Individual ratings were correlated with each item parameter to investigate individual rater sensitivity to the item parameter characteristics.
- Results suggested that raters were not sensitive to *a*- and *c*-*parameters.*
- However, several raters were attentive to *b-parameters* in the data set - negative correlations suggested that these raters assigned lower *p-value* estimates as item difficulty increased.

Conclusions

- Fundamental question - How do discrimination indices relate to variance among Angoff judgments?
- No significant relationships were found with the current data set due to non-rater fit.
- Implications:
  - Assist test developers to identify the most accurate Angoff raters for panels.
  - Analysis by means of IRT can aid researchers in uncovering influences conflicting with the consensual MQC definition (non-rater fit).
  - Provide further evidence to improve rater training and discussion of the MQC.
  - Isolate specific item parameters to which raters are not anchored and develop training to encourage proper adjustment.

Conclusions

- Supported Kane and Van der Linden's conclusions
  - Inappropriate MQC anchoring or non-rater fit will confound data.
- Larger question exists:
  - If Angoff raters do not attend to item characteristics properly, how reliable are the rater judgments in establishing cutoff scores?
  - How reliable is the Angoff method in establishing cutoff scores?
  - Raters who cannot properly estimate the height or steepness of the ICC at the MQC anchor will select a pass point that will not reflect an honest demarcation of ability.

Thank You!

chrisj@cps.ca.gov
www.cps.ca.gov

CPS
Human Resource Services