

Ch. 6: Test Development

748

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Copyright © 2022 Michael Diehr
All Rights Reserved
For use only by students enrolled
in my sections of Psyc 402
through the end of the semester.
May not be posted, shared or uploaded
online without permission.

749

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Review

751

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Ch. 6: Test Development

- Test Items
 - question formats (T/F, Multiple Choice, Likert...)
- Guessing & Correction for guessing formula
- Cognitive Factors: Recall vs. Recognition
- Exercise: from construct to question
- Item Analysis: Difficulty, Discriminability, ICC
- Item Response Theory / Adaptive Testing
- SII (Strong Interest Inventory)

755

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Writing test items

- Define what you are measuring (theory of the construct)
- Write many items that cover the *content*
- Avoid very long items
- Use appropriate reading level
- Don't mix two concepts in one question.
- Vary the "response set" with both positively and negatively worded items

756

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Test Item Formats

- Qualitative
 - Fill in the blank
 - Essay
- Quantitative
 - True / False...
 - Multiple Choice...
 - Rating / Category scales...

757

Psychology 402 - Fall 2022 - Dr. Michael Diehr

Dichotomous Format

- Aka “True/False” or “Yes/No” or “Binary”
- Pros: easy to write, administer, and score, good for basic facts. Avoids ambivalence.
- Cons: rote memorization, high scores due to guessing → increased # of items, black & white thinking: not appropriate for complexity or nuance
- Summary: unsophisticated format - shouldn't be widely used for achievement testing

758

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Poly[cho]tomous

- AKA “multiple choice”
- Target: correct answer
- Distractor: incorrect answers
- Pros: easy to administer (covers a lot of material quickly), easy to score, can handle shades of gray / nuance
- Cons: difficult to write, susceptible to guessing strategies, susceptible to “over studying”

759

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Miller's Law

- To fully characterize a Normal distribution of X, we need its Mean and ...?
 - A: Mode
 - B: X-bar
 - C: Standard Deviation
 - D: Average of X

760

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Distractors?

- Too few distractors --> dichotomous
- Too many distractors --> slow, confusing
- Optimal is 3-5 distractors. Thus, most multiple-choice tests should have between 4 and 6 possible answers per question.
- Distractors should cover a wide range of abilities w/o being cute or trite

761

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Guessing : Probability

- M = # of answer choices per question
- P_{correct} with random guessing = $1/M$
- On a dichotomous (T/F), $P = \underline{\hspace{1cm}}$
- On a multiple choice test with M answers per question, the probability = $\underline{\hspace{1cm}}$
- Total score from guessing:
 - $N_{\text{questions}} \times P_{\text{correct}}$

763

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Guessing : Expected Score

- Probability of getting any item correct, using a random guessing strategy, p is equal to 1 divided by the # of answers.
- On a dichotomous (T/F) test the probability $P = 1/2 = 50\% = 0.5$
- On a multiple choice test with M answers per question, the probability = $1/M$. For a 4 item test $P = 1/4 = .25 = 25\%$
- Total score due to guessing = # of questions times average score per item or $N \times P$.
- Example: an 100 item test with 4 answers = 25

764

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Guessing impacts Validity

765

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Correcting for Guessing

- Scores can correct for guessing.
- Goal: person randomly answering should get same score as someone who doesn't answer.
- Expected score of someone who answers no questions = 0
- Expected score of someone who guesses randomly is $N * (1/M)$
- Correction Formula:
 - For every wrong answer, subtract $1/(M-1)$ points.

766

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Correcting for Guessing : Example

- Example:
 - a 100 item test ($N=100$)
 - each question has 5 choices ($M=5$)
 - probability of right answer by guess? ($P = 1/M = 1/5 = 20\%$)
- A student guessing on each item would average 20 correct ($P*N = 0.2 * 100 = 20$)
- Correction: subtract $(1/M-1)$ points for each wrong answer = $1/(5-1) = 1/4 = 0.25$ points.
- Adjusted score?

767

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Correcting for Guessing - Real World

- Formula is simplistic
- College Board removed guessing penalty for AP exams in 2010
- SAT revisions in 2016
 - Removes penalty for Guessing
 - other changes:
 - Essay is optional
 - Vocabulary test changed

768

Psychology 402 - Fall 2022 - Dr. Michael Dohr

When should you guess?

- Almost always
- Worst case: if a correction formula is in use, and you truly have zero information for a given item, guessing has no effect
- However, it's likely you do have some knowledge. This increases your chances slightly above chance, giving you a positive expected score.

769

Psychology 402 - Fall 2022 - Dr. Michael Dohr

[di | poly]chotomous Issues

- Pros:
 - neutral, fair scoring
- Types of knowledge:
 - Recall vs. Recognition
 - Receptive vs. Expressive
- Skill =? test taking ability
- Solution: Essay test format

770

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Accessing Knowledge

- Recalling information is different than Recognizing it
- Neuropsychology suggests different brain systems. Recall can be stronger or weaker than Recognition
- Issues for testing:
 - What type of access is involved in polychotomous testing?
 - Is it fair to test using a method which prefers one type over the other?

771

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Recall vs. Recognition

772

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Facts vs Opinions?

- Polychotomous : good for assessing factual information
- What about measuring opinions, preferences, styles?

777

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Other question formats

- Likert Scale
- Category Rating Scale
- Visual Analogue Scale
- Checklists

778

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Likert Format

- Asked to rate statements on an ordinal scale with a short list of answer choices
- Example:
I am afraid of heights:
 - 1 strongly disagree
 - 2 disagree
 - 3 undecided
 - 4 agree
 - 5 strongly agree
- Numbers : sometimes shown, sometimes not shown.

780

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Likert : Neutral?

- Sometimes, want to avoid the middle (neutral, undecided) answer
- Example:
I am afraid of heights:
 - 1 strongly disagree
 - 2 somewhat disagree
 - 3 somewhat agree
 - 4 strongly agree
- Like T/F, forces subject to take a position

781

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Likert : Balance & Symmetry

- Answers should be balanced & symmetrical
- Example:
 - 1 strongly disagree
 - 2 somewhat disagree
 - 3 neutral
 - 4 somewhat agree
- Poor design
 - Answers will be biased towards 3 or 4

782

Psychology 402 - Fall 2022 - Dr. Michael Dohr

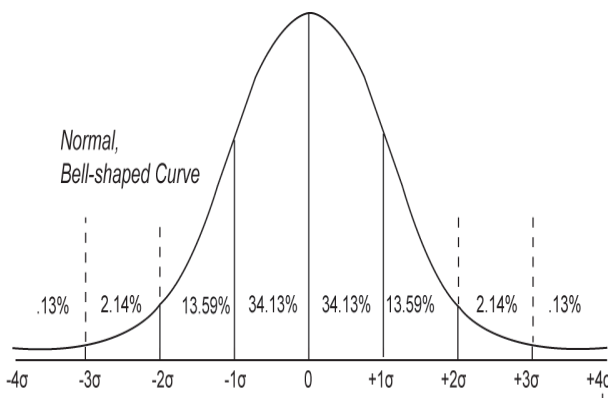
Likert Scales : 6 and 7 choices

- | | |
|----------------------|----------------------|
| 1. Strongly Disagree | 1. Strongly Disagree |
| 2. Disagree | 2. Disagree |
| 3. Somewhat Disagree | 3. Somewhat Disagree |
| 4. Somewhat Agree | 4. Neutral |
| 5. Agree | 5. Somewhat Agree |
| 6. Strongly Agree | 6. Agree |
| | 7. Strongly Agree |

783

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Ideal # of answers?



Psychology 402 - Fall 2022 - Dr. Michael Dohr

Category (Rating Scale) Format

- Similar to Likert format, but #s are used instead
- Pros -- responses are more precise than with Likert scales (10 vs. 5 or 6)
- Cons -- context effects stronger
 - Solution: clearly define endpoints
- Precision vs. Accuracy?

785

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Rating Scale - no anchors

- On a 1 to 10 scale how much do you like your partner?
 - 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
- Issues:
 - Is 1 or 10 the top?

786

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Rating Scale - with anchors

- On a 1 to 10 scale how much do you like your partner?
 - 1 Planning to break up
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10 Planning to get Married soon
- Issues:
 - Unbalanced (is 5 or 6 the middle?)
 - Interpretation? what does a "2" or "3" mean?

787

Psychology 402 - Fall 2022 - Dr. Michael Dohr

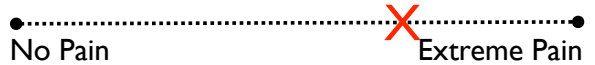
How many choices?

- Optimal # of choices is between 4 and 7
 - consistent with Miller's 7 ± 2
- Using up to 10 choices is OK if
 - good anchors & examples are provided
 - raters are motivated or trained
 - Otherwise, using 10 choices leads to random responding

788

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Visual Analogue Scale

- Similar to Category format, except use of a visual stimulus & graphical measurement
- Example:
How much pain are you in right now?

- Pros: allows a precise, finely detailed response
Cons: hard to score, precision vs. accuracy?

789

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Checklists

- Checklists:
 - Agree/disagree with large # of statements
- Example
 - “I am currently having trouble with...”
 - ☐ Money
 - ☐ Relationships
 - ☐ Appetite
 - ☐ Sleep
 - ☐ ...

792

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Advice from Textbooks

| Advice | % endorsing |
|---|-------------|
| Don't use “All of the above” | 80% |
| Don't use “None of the Above” | 75% |
| All choices should be plausible | 70% |
| Negative wording shouldn't not be un-used | 55% |

793

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Exercise: From Construct to Question

794

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Ch. 6 - Part 2

799

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Review

- Reliability and Validity of entire Test
- Individual Test Items
 - dichotomous / polychotomous
 - recall vs. recognition
 - Likert
 - neutral, balanced
 - Category
 - anchors, context effects
- Ideal # of answers per question?

801

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Item Analysis

- In Ch 5 we discussed the reliability and validity *of the entire test*.
- Now we look at psychometrics of *individual test items*.
- Item Difficulty
- Item Discriminability

803

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Item Difficulty

- How hard is this item?
- % who get the item correct (item easiness)

804

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Too hard / Too easy

- Floor effect: many scores near the bottom range of possible scores
- Ceiling effect: many scores near the top range of possible scores

805

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Ideal Difficulty

- Ideal= halfway between chance and perfect
 - for a 4-item multiple choice, chance = 25%, so optimum would be 62.5%
 - typical range is 30% to 70%
- Tests should contain wide variety of item difficulties, because people are different

806

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Ideal Difficulty 2

- Mathematically, 30%-70% is optimum
- What about human / emotional issues?
 - Tests or items that are too hard?
 - Tests or items that are too easy?

807

Psychology 402 - Fall 2022 - Dr. Michael Diefel

Discriminability

- Difficulty = how many people answer correctly?
- Discriminability = who answers correctly?
- Does performance on one item correlate with overall test performance?
- Two ways
 - statistical
 - graphical

808

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Discriminability - Statistical

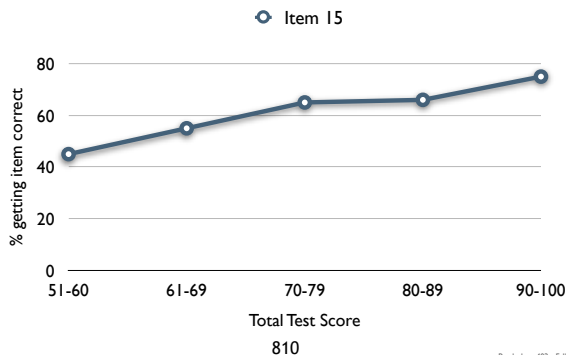
- Extreme Group:
 - divide test takers into thirds
 - % correct : top third vs. bottom third
- Point Biserial
 - p.b. correlation between item and test score
 - low or negative values represent “bad” items

809

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Discriminability - Graphical

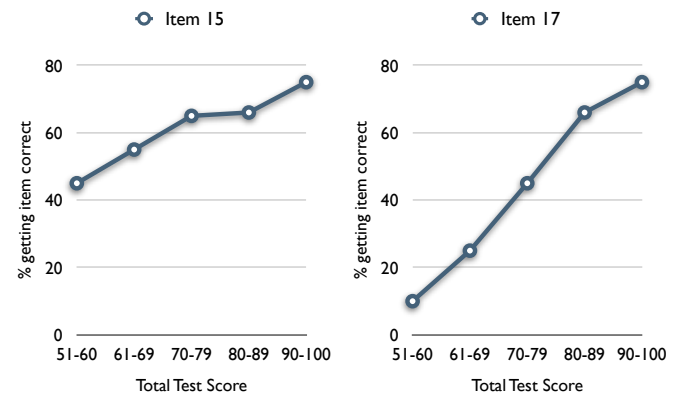
- Item Characteristic Curve
- Graph % correct vs. total test score for one test item



Psychology 402 - Fall 2022 - Dr. Michael Dohr

Item Characteristic Curve

- Different test items have different ICCs

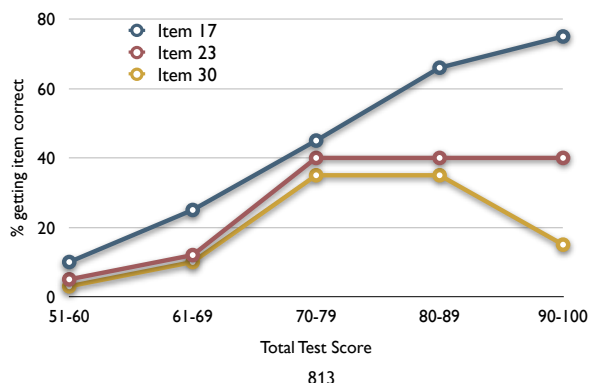


812

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Item Characteristic Curve

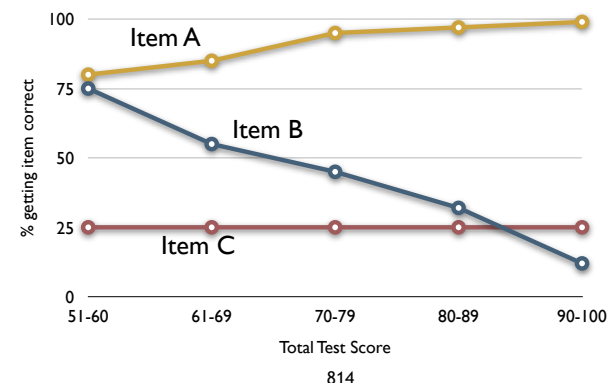
- Good items show steady increase
- Bad items show decreases or flat spots



Psychology 402 - Fall 2022 - Dr. Michael Dohr

ICC Example

- Diagnose these problems:



814

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Graph the ICC

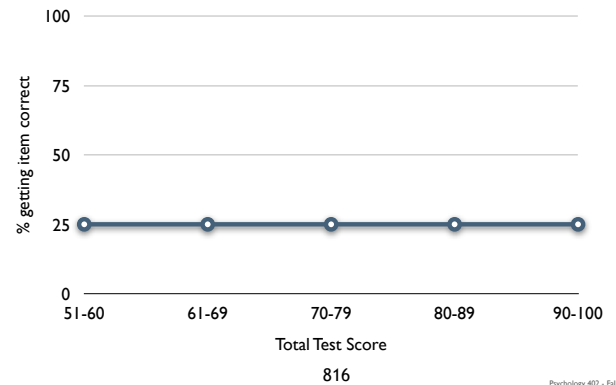
- Item 1: What was the exact population of the town Bodie, California, in 1879?
(A) 6142
(B) 6143
(C) 6144
(D) 6145
- Correct answer = A

815

Psychology 402 - Fall 2022 - Dr. Michael Diehr

ICC Example

- Random guessing



Psychology 402 - Fall 2022 - Dr. Michael Diehr

Graph the ICC

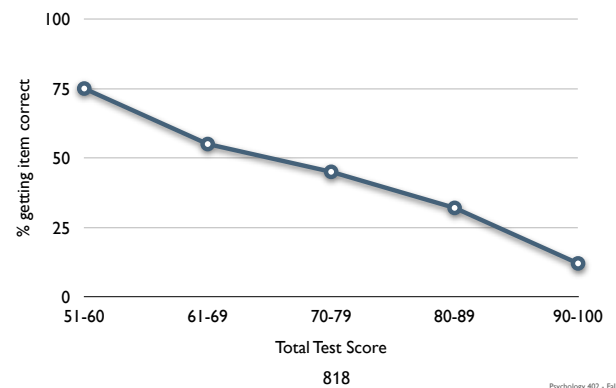
- Item 1: What is 0.34 times 0.27
(A) 9.18
(B) 0.61
(C) 0.0918
(D) 91.8
- “Correct Answer” = B

817

Psychology 402 - Fall 2022 - Dr. Michael Diehr

ICC Example

- Answer key is wrong



Psychology 402 - Fall 2022 - Dr. Michael Diehr

Graph the ICC

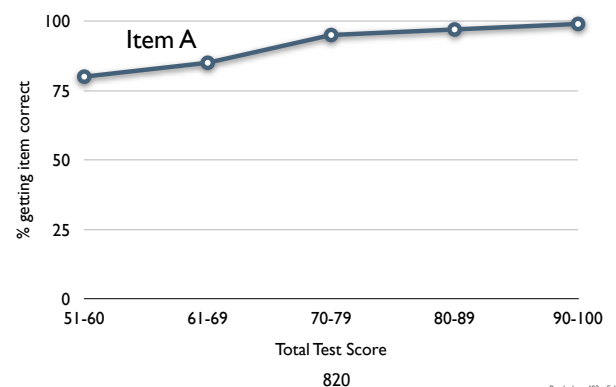
- Item 1: What is $1 + 2$
(A) 11
(B) 21
(C) 3
(D) 0.3
- Correct answer = C

819

Psychology 402 - Fall 2022 - Dr. Michael Diehr

ICC Example

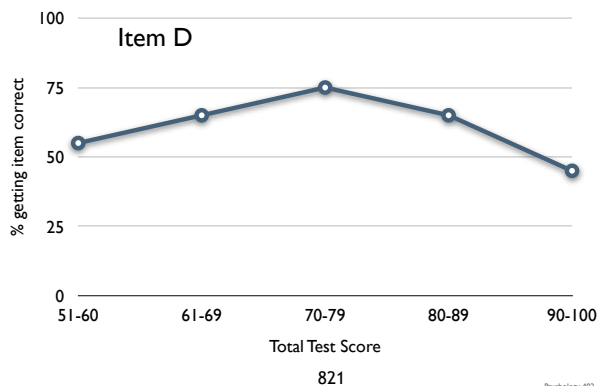
- Item is too easy



Psychology 402 - Fall 2022 - Dr. Michael Diehr

ICC Example

- The “Overstudying” problem:



Psychology 402 - Fall 2022 - Dr. Michael Dohr

Q: How many Human Genders are there?

- A : One (Human)
- B : Two (Male, Female)
- C : Three (Male, Female, Neuter)
- D : Four (Male Adult, Male Child, Female Adult, Female Child)
- E : None of the above

822

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Item Response Theory (IRT)

- Classical Test theory
 - your ability = *number of items correct*
- IRT
 - your ability = *level of difficulty* at which you can perform
- IRT Model : probability of correct answer is modeled using several variables (for the test and the test-taker)
- IRT Procedures: computer-based *adaptive testing*

823

Psychology 402 - Fall 2022 - Dr. Michael Dohr

IRT / Adaptive Testing

- To cover different ability levels, tests need wide range of item difficulties
- For an individual, some items will be too easy / some too hard
- “old fashioned” solution = have several tests (easy...medium...hard) and pick a test based on pre-existing knowledge of person.
- IRT solution = one test that automatically detects person’s level and gives questions mainly in that difficulty level.

824

Psychology 402 - Fall 2022 - Dr. Michael Dohr

IRT in the real world

- IRT is theoretically better
- Adoption in curriculum is slow
- some tests use it but vast majority do not
- Continuing research

825

Psychology 402 - Fall 2022 - Dr. Michael Dohr

External Criteria

- Internal Criteria = total test score
- External Criteria = thing that actually matters (e.g. “do you crash the plane”)
- Most Item Analysis still uses Internal criteria rather than the more correct External Criteria
- Why?

826

Psychology 402 - Fall 2022 - Dr. Michael Dohr

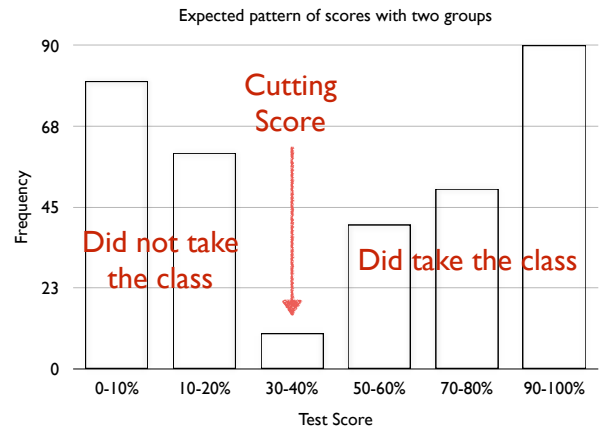
Criterion-referenced Test

- Instead of arbitrary criteria such as “70% = pass” use one with more validity.
- Criteria = the learning outcome(s) desired
- Method:
 - create a good test
 - give it to two groups of students
 - those who have had the material
 - those who have not
 - Determine cut-point score from histogram

827

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Criterion-referenced Test



828

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Criterion-referenced Test

829

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Limitations of Item Analysis

- Tests discriminate between levels of performance
- Statistics (difficulty and discriminability) don't tell why a person missed an item
- Items might discriminate well (statistically) but for the wrong reasons (educationally)
- Tests don't directly help people learn
- Tests can harm, if they dramatically change learning behavior (e.g. study for the test rather than the subject)

830

Psychology 402 - Fall 2022 - Dr. Michael Dohr

Example of a poor test item?

- What is 0.4 plus 0.3
 - (A) 0.3
 - (B) 0.4
 - (C) 0.7
 - (D) .07
- Is answering (A) better or worse than answering (D)?

831

Psychology 402 - Fall 2022 - Dr. Michael Dohr