Ch. 6: Test Development

Ch. 6: Test Development

- Goals of this chapter
- Test Items
  - Common formats
  - Alternative formats
- Item Analysis
  - Item Difficulty
  - Discriminability
- Item Response Theory

Ch. 6: Goals

- Understand several test item formats
- Correction for guessing on Multiple Choice exams
- Understand rating scales (Likert, 10 point, etc.)
- Measure and adjust item Difficulty
- Measure and adjust item Discriminability
- Item Characteristic Curves
- Describe the “over studying” problem
- Limitations of item analysis / item response theory.

Test Item Formats

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

Writing test items...

- Define what you are measuring using a theoretical framework (the “Construct”)
- Write a large pool of items that cover the *content* area without duplication
- Avoid very long items
- Use a reading level difficulty appropriate for the test takers
- Avoid complexity -- don’t mix two concepts in one question.
- Vary the “response set” with both positively and negatively worded items

Dichotomous Format

- Aka “True/False” or “Yes/No” test
- Pros: easy to write, administer, and score, appropriate for simple facts. Avoids ambivalence.
- Cons: rote memorization, high scores due to guessing —> increased # of items, punishes complexity or nuanced thinking, black & white thinking: not appropriate for complexity

- Summary: unsophisticated format that should not be widely used for achievement testing

Poly[cho]tomous

- AKA “multiple choice”
- Target: correct answer
- Distractor: incorrect answers
- Pros: easy to administer (cover a lot of material quickly vs essay test), easy to score, can handle shades of gray or nuances of meaning
- Cons: difficult to write, susceptible to guessing strategies, susceptible to “over studying”

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

Guessing : Probability

- Probability of getting any item correct, using a random guessing strategy, is equal to 1 divided by the # of answers.
- On a dichotomous (T/F) test the probability = ____
- On a multiple choice test with M answers per question, the probability = ______

- Total score due to guessing = # of questions times average score per item or ________

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 * P.
- Example: an 100 item test with 4 answers = 25

Correcting for Guessing

- Scores can correct for guessing.
- Goal is to equalize the scores of someone who guesses randomly with someone who doesn’t answer
- Expected score of someone who answers no question = zero
- Expected score of someone who guesses randomly is N* (1/M)
- For every wrong answer, subtract 1/(M-1) points.

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 takes the test, guesses on each item, and gets 20 correct (P*N = 0.2 * 100 = 20)
- Correction for guessing subtracts (1/M-1) points for each wrong answer = 1/(5-1) = 1/4 = 0.25 points.
- Adjusted score?

Correcting for Guessing - Real World

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

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 gains you nothing
- However, chances are that you actually have some knowledge. This increases your chances slightly above chance, giving you a positive expected score.

[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

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?

Recall vs. Recognition

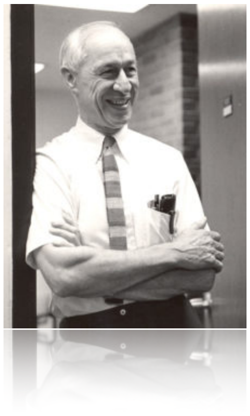
Other question formats

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

Rensis Likert

American social psychologist

Pronounced “LICK-ert”



Likert Format

- Asked to rate statements on a scale with a small fixed number of answers
- Example:
  I am afraid of heights:
  1 strongly disagree
  2 disagree
  3 undecided
  4 agree
  5 strongly agree
- Numbers : sometimes shown, sometimes not shown.

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

Likert : Balance & Symmetry

- Answers should be balanced & symmetrical in all cases
- Example:
- I am afraid of heights:
  1 strongly disagree
  2 somewhat disagree
  3 neutral
  4 somewhat agree
- Poor design
  - Answers will be biased towards 3 or 4

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?

Category Example

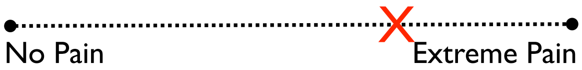
- On a 1 to 10 scale how much do you like your partner?
  1 Planning to break up
  2
  3
  4
  5
  6
  7
  8
  9
  10 Planning to get Married soon
- Issues:
  - Unbalanced (is 5 or 6 the middle?)
  - Hard to interpret : what does a “2” or “3” really mean?

How many choices?

- Research suggests optimal # of choices is between 4 and 7
- Using up to 10 choices is OK if
  - raters are motivated
  - good anchors & examples are giving
  - Otherwise, 10 choices leads to random responding

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?

Checklists

- Checklists:
  - Agree/disagree with large # of statements
- Example

- "I am currently having trouble with…"

□ Money

□ Relationships

□ Appetite

□ Sleep

□ …

Q sorts

- Q sort:
  - sort large # of statements into piles depending on how much you agree/disagree (like Likert format)
  - Responses follow bell-shaped curve, extreme responses are most interesting

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%

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

Item Difficulty

- How hard is this item?
- % who get the item correct (item easiness)
- Ideal Difficulty is halfway between chance-level performance and 100%
  - e.g. for a 4-item multiple choice, chance = 25%, so optimum would be 62.5%
  - typical range is 30% to 70%
- Test as a whole should have wide variety of item difficulty in order to work with diverse subjects.

Item 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?

Discriminability

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

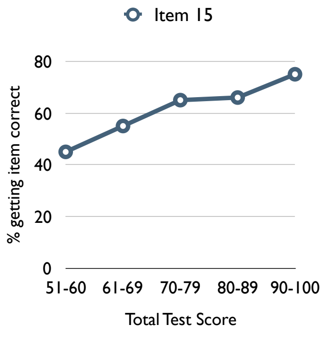
- Two ways
  - statistical
  - graphical

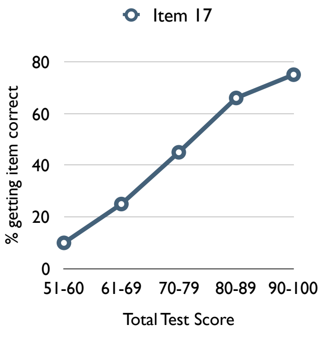
Discriminability - Statistical

- 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

Discriminability - Graphical

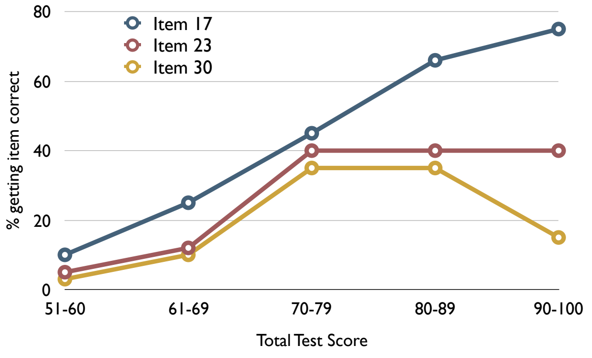
- Item Characteristic Curve
- Graph % correct vs. total test score for one test item





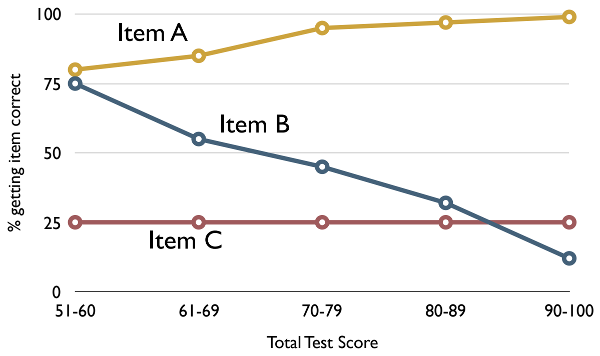
Item Characteristic Curve

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



ICC Example

- Diagnose these problems:



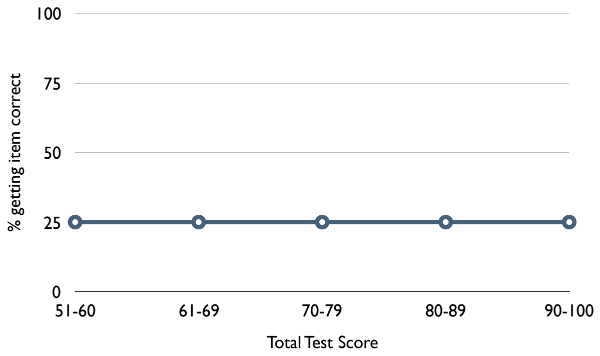
Graph the ICC

- 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

ICC Example

- Random guessing



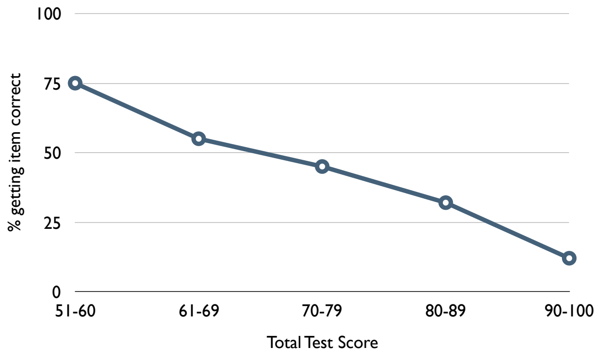
Graph the ICC

- Item 1: What is 0.34 times 0.27
  (A) 9.18
  (B) 0.61
  (C) 0.0918
  (D) 91.8

- “Correct Answer” = B

ICC Example

- Test item has wrong answer



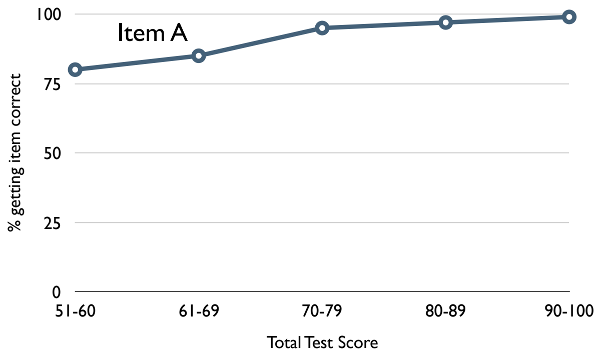
Graph the ICC

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

- Correct answer = C

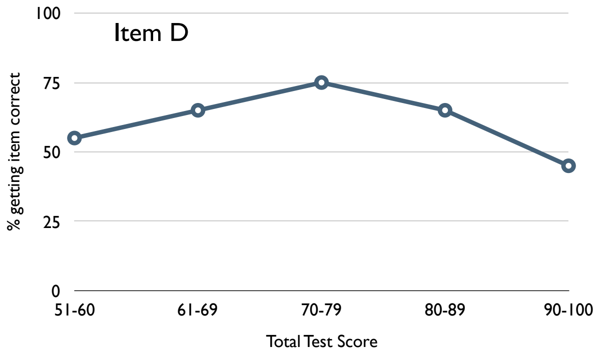
ICC Example

- Item is too easy



ICC Example

- “Overstudying” or “None of the above



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

Item Response Theory (IRT)

- Classical Test theory : score = # of items correct
- IRT: score = level of difficulty at which you can perform
- IRT Model : probability of correct answeris modeled using formal parameters (of the Person and the Test)
- IRT Procedures: using computer-based adaptive testing, test questions are given to focus in on the ability level of the test subject

IRT / Adaptive Testing

- To cover a range of ability levels, tests must have a range of item difficulties
- For a person (who has one ability level) many items are too easy and many too hard.
- “old fashioned” solution = have many tests, choose right one 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.

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

External Criteria

- Internal Criteria = total test score
- External Criteria = thing that actually matters (e.g. “do you crash the plane”)

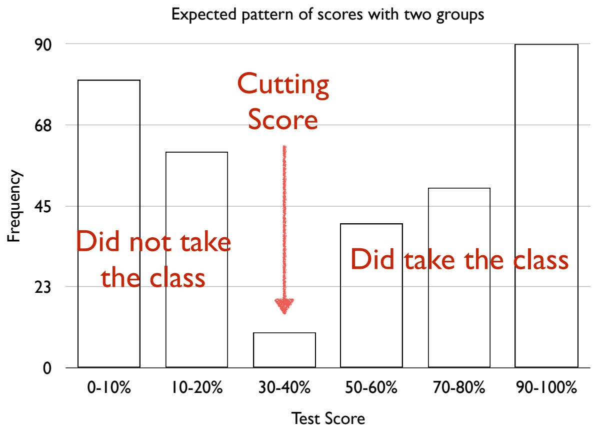
- Most Item Analysis still uses Internal criteria rather than the more correct External Criteria
- Why?

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

Criterion-referenced Test



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)

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)?

Strong Interest Inventory (SII)

The Structure of the SII

Section 1 : General Occupational Themes
Section 2 : Basic Interest Scales
Section 3 : Occupational Scales
Section 4 : Personal Style Scales
Section 5 : Profile Summary
Section 6 : Response Summary

About the SII

- 291 multiple choice questions (polychotomous)
- Likert-style questions
- Takes about 25 minutes to take
- Developed in 1927 by E.K. Strong, Jr.
- Vocational placement upon leaving military
- Based partly on “Holland Codes”

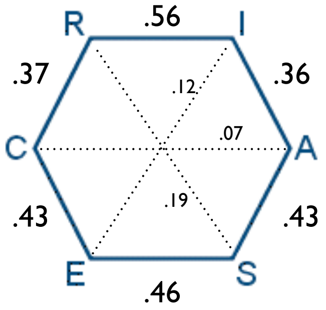
Holland Typology

- Theory: personality and vocations share six main *factors*

Type	Description
Realistic	practical, physical, hands-on, tool-oriented
Investigative	analytical, intellectual, scientific, explorative
Artistic	creative, original, independent, chaotic
Social	cooperative, supporting, helping, healing/ nurturing
Enterprising	competitive environments, leadership, persuading
Conventional	detail-oriented, organizing, clerical

Holland Typology

- Type : usually expressed as top 3 factors
- Hexagon indicates correlation between factors

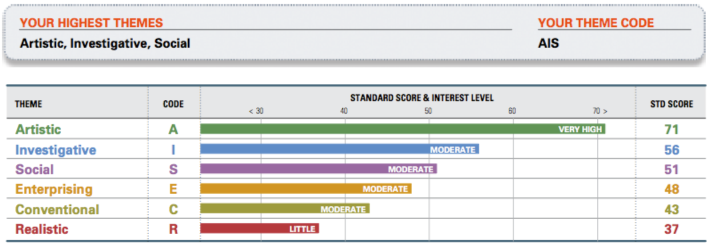


SII uses T-Scores

	Z scores	IQ scores	T scores	Scaled Scores
Mean	0	100	**50**	10
SD	1	15	**10**	3

1 : General Occupational Themes (GIS)

Describes your interests, work activities, potential skills, and personal values in six broad areas: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C).



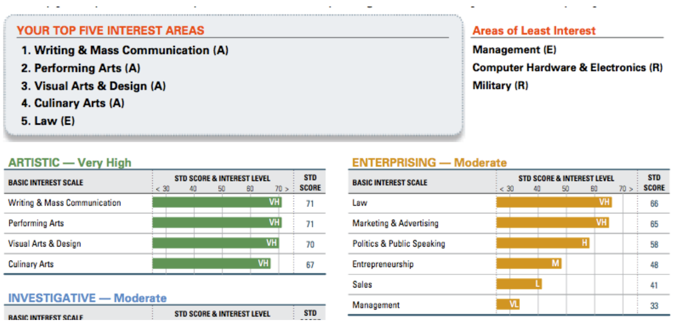
YOUR HIGHEST THEMES
Artistic, Investigative, Social

YOUR THEME CODE
AIS

THEME	CODE	STANDARD SCORE & INTEREST LEVEL	STD SCORE
Artistic	A	VERY HIGH	71
Investigative	I	MODERATE	56
Social	S	MODERATE	51
Enterprising	E	MODERATE	48
Conventional	C	MODERATE	43
Realistic	R	LITTLE	37

2: Basic Interest Scales (BIS)

Identifies specific interest areas within the six General Occupational Themes, indicating areas likely to be most motivating and rewarding for you.



YOUR TOP FIVE INTEREST AREAS

1. Writing & Mass Communication (A)
2. Performing Arts (A)
3. Visual Arts & Design (A)
4. Culinary Arts (A)
5. Law (E)

Areas of Least Interest

Management (E)
Computer Hardware & Electronics (R)
Military (R)

ARTISTIC — Very High

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30, 40, 50, 60, 70 >)	STD SCORE
Writing & Mass Communication	VH	71
Performing Arts	VH	71
Visual Arts & Design	VH	70
Culinary Arts	VH	67

ENTERPRISING — Moderate

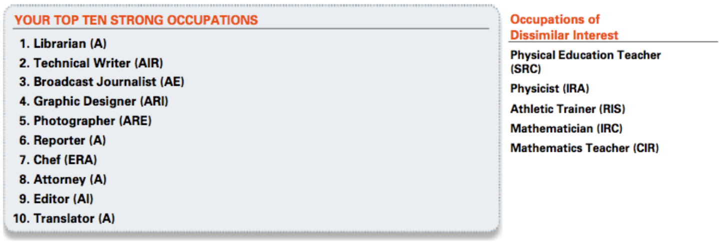
BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL (< 30, 40, 50, 60, 70 >)	STD SCORE
Law	VH	66
Marketing & Advertising	VH	65
Politics & Public Speaking	H	58
Entrepreneurship	M	48
Sales	L	41
Management	VL	33

INVESTIGATIVE — Moderate

BASIC INTEREST SCALE	STD SCORE & INTEREST LEVEL	STD

3 : Occupational Scales (OS)

Compares your likes and dislikes with those of people who are satisfied working in various occupations, indicating your likely compatibility of interests.



YOUR TOP TEN STRONG OCCUPATIONS

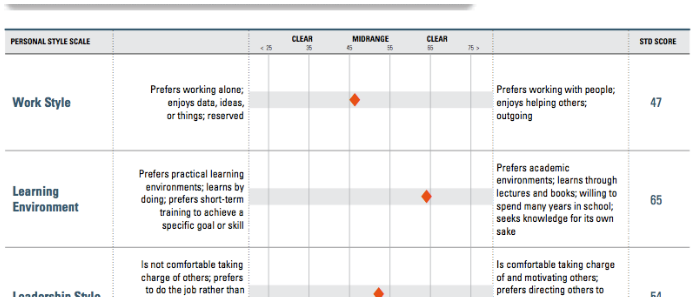
1. Librarian (A)
2. Technical Writer (AIR)
3. Broadcast Journalist (AE)
4. Graphic Designer (ARI)
5. Photographer (ARE)
6. Reporter (A)
7. Chef (ERA)
8. Attorney (A)
9. Editor (AI)
10. Translator (A)

Occupations of Dissimilar Interest

Physical Education Teacher (SRC)
Physicist (IRA)
Athletic Trainer (RIS)
Mathematician (IRC)
Mathematics Teacher (CIR)

4 : Personal Style Scales (PSS)

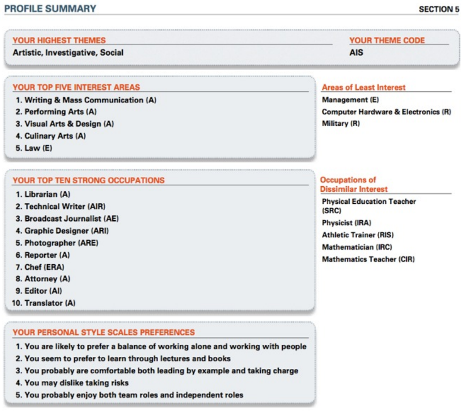
Describes preferences related to work style, learning, leadership, risk taking, and teamwork, providing insight into work and education environments most likely to fit you best.



PERSONAL STYLE SCALE		CLEAR < 25 35 · MIDRANGE 45 55 · CLEAR 65 75 >		STD SCORE
Work Style	Prefers working alone; enjoys data, ideas, or things; reserved	◆	Prefers working with people; enjoys helping others; outgoing	47
Learning Environment	Prefers practical learning environments; learns by doing; prefers short-term training to achieve a specific goal or skill	◆	Prefers academic environments; learns through lectures and books; willing to spend many years in school; seeks knowledge for its own sake	65
Leadership Style	Is not comfortable taking charge of others; prefers to do the job rather than	◆	Is comfortable taking charge of and motivating others; prefers directing others to	54

5 : Profile Summary

Provides a graphic snapshot of Profile results for immediate, easy reference.



PROFILE SUMMARY SECTION 5

YOUR HIGHEST THEMES
Artistic, Investigative, Social

YOUR THEME CODE
AIS

YOUR TOP FIVE INTEREST AREAS

1. Writing & Mass Communication (A)
2. Performing Arts (A)
3. Visual Arts & Design (A)
4. Culinary Arts (A)
5. Law (E)

Areas of Least Interest

Management (E)
Computer Hardware & Electronics (R)
Military (R)

YOUR TOP TEN STRONG OCCUPATIONS

1. Librarian (A)
2. Technical Writer (AIR)
3. Broadcast Journalist (AE)
4. Graphic Designer (ARI)
5. Photographer (ARE)
6. Reporter (A)
7. Chef (ERA)
8. Attorney (A)
9. Editor (AI)
10. Translator (A)

Occupations of Dissimilar Interest

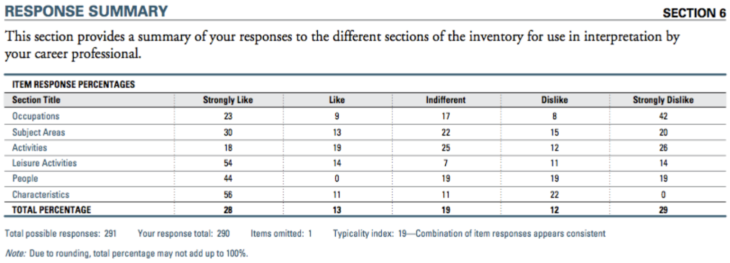
Physical Education Teacher (SRC)
Physicist (IRA)
Athletic Trainer (RIS)
Mathematician (IRC)
Mathematics Teacher (CIR)

YOUR PERSONAL STYLE SCALES PREFERENCES

1. You are likely to prefer a balance of working alone and working with people
2. You seem to prefer to learn through lectures and books
3. You probably are comfortable both leading by example and taking charge
4. You may dislike taking risks
5. You probably enjoy both team roles and independent roles

6 : Response Summary

Summarizes your responses within each category of Strong items, providing interpretive data useful to your career professional.



RESPONSE SUMMARY SECTION 6

This section provides a summary of your responses to the different sections of the inventory for use in interpretation by your career professional.

ITEM RESPONSE PERCENTAGES

Section Title	Strongly Like	Like	Indifferent	Dislike	Strongly Dislike
Occupations	23	9	17	8	42
Subject Areas	30	13	22	15	20
Activities	18	19	25	12	26
Leisure Activities	54	14	7	11	14
People	44	0	19	19	19
Characteristics	56	11	11	22	0
TOTAL PERCENTAGE	**28**	**13**	**19**	**12**	**29**

Total possible responses: 291 Your response total: 290 Items omitted: 1 Typicality index: 19—Combination of item responses appears consistent

Note: Due to rounding, total percentage may not add up to 100%.

SII Reliability

- Generally good Reliability

Type	Cronbach's Alpha	Test-Retest
GOTs	.91 - .95	.84 - .92
BIS	0.87	
Occupational Scales		.82 - .89

SII Validity

- Concurrent Validity
  - measured % Hit Rate for using Occupational Scale to predict College Major
  - Excellent or Moderate hit:
    - 82% for females, 92% men
- Predictive Validity
  - % hit rate for major Senior for tests taken as Freshmen (3.5 years)
  - 69% females,70% for males

Career Paths in Psychology

- This is an optional discussion