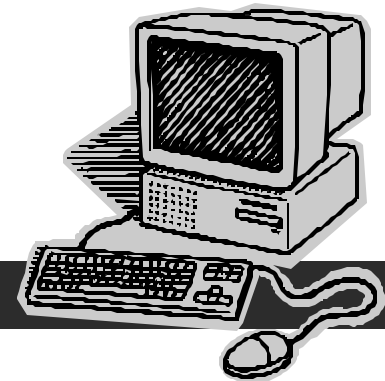


# **The Art & Science of Developing Hybrid Situational Judgment and Knowledge/Ability Based Exams** *Or How I Grew to Love Computerized Testing*

Presented by  
Bobbie Ames  
James Frankart



Commonwealth of Pennsylvania

# What is a hybrid exam?

- A computerized exam that combines two or more methods of measurement within a single exam to assess critical job requirements



# In The Beginning...

- 1999 arbitration agreement to pilot a promotional exam
- Written test required
- Test material subject to approval of union
- No previous examination
- ICE implementation was pending, then deferred

# **Game Conservation Officer Supervisor Promotional Exam**

The Developmental  
Process





# Pennsylvania Game Commission

- Pennsylvania's wildlife management agency
- 700 full-time employees and several thousand part-time employees
- Central HQ in Harrisburg
- Six regional offices with five program areas
- Supervisors promoted without exam from Conservation Officers and Land Management Group Supervisors job classes

# Game Conservation Officer Supervisors

- Minimum Qualifications: “Three years of experience in either Land Management or Game Protection as a Game Conservation Officer with the Pennsylvania Game Commission”

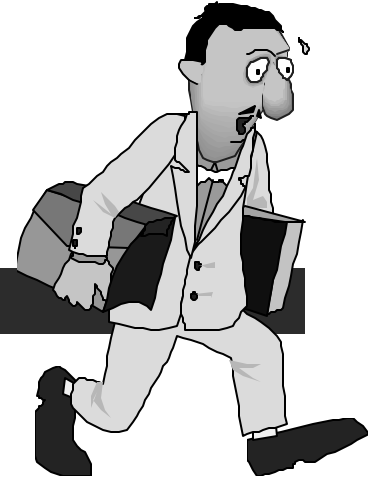


# The Job Study - Considerations

- Diverse jobs – 5 program areas
- 29 incumbents
- Six regional offices
- Regional differences
- Personnel Director's advice & assistance
- Exceptional cooperation from everyone!



# SMEs



- Six PGC Regional Directors
- Developed and operationally defined WBs and KSAs
- Rated WBs and KSAs, determined linkage, and consolidated KSAs to four factors for test purposes

# Establishing Content Validity

- Interviewed 15 and observed 8 incumbents
- Surveyed all incumbents for WBs & KSAs and ratings
- Interviewed Bureau Directors for all program areas



# Example - Work Behaviors

- Administers and plans the daily operations of the Pennsylvania Game Commission (PGC) programs that promote habitat development and management, wildlife management, conservation education, public relations, and game law enforcement by assessing program needs, developing resources, deciding how resources will be used, and planning for future needs

## Example - Work Behaviors (cont'd)

- Reviews, compiles and/or prepares reports concerning various activities such as training, meetings, projects, prosecutions, expenses, complaints, equipment usage, personnel actions, wildlife surveys, and program implementation



# Examples - KSAs

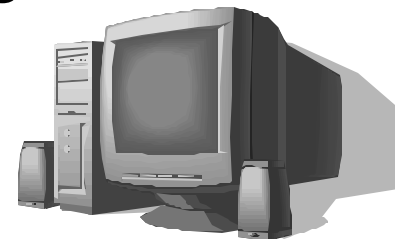
- Knowledge of the programs and practices of the Pennsylvania Game Commission
- Knowledge of wildlife identification and habitat management practices and methods
- Ability to exercise proper judgment in various situations encountered on the job including those of a serious or unusual nature
- Ability to communicate effectively in writing

# Identifying Job Requirements

- Ratings made by incumbents, supervisors, and bureau directors (2nd line supervisors) to identify the important, entry-level job requirements
- The KSAs rated by > 80% as entry-level and received a combined rating of 2.5 on a scale of 1 to 3 in overall importance and in RSP were selected to be measured

# Test Planning Considerations

- Homogeneous candidate group
- Estimate 120 eligible candidates
- Test security concerns
- Computer issues
- No oral exam
- Times & locations for testing



# Test Design & Development

- Entry-level KSAs categorized into four factors to be measured
- Operationally defined each factor
- Discussed and determined most effective method of measurement for each factor
- Considered resources available, validity of method, ease of administration, scope of concept

# Measuring Technical Job Knowledge

- Measured with standard multiple choice questions to determine knowledge of facts
- High reliability, easy to machine score, but can be difficult to write good questions
- No database items were suitable to SMEs so many were developed and refined
- 45 new TJK items were ultimately used

# Measuring Effective Working Relationships

- Measured by method applying the ability to evaluate a situation and determine what to do and say
- Situational Judgment Test selected – at least moderate reliability, some evidence that similar candidate experience and real-life job-related situations increase validity as a measurement method

# Measuring Judgment

- Also selected to measure with SJT
- Acting effectively based on knowing facts, assessing a situation, and evaluating choices while considering the implications of problems and solutions
- Ruled out in-basket exercise which may have measured more static information concerning procedures for this job instead of a dynamic unfolding of a demanding situation

# Measuring Written Communication

- Writing exercise is valid method to measure what candidate chooses to say, organizes it, and then presents it
- Face validity increases candidate acceptance
- Reliability can be questionable because of human rater tendency to subjectivity and time involved. Can control somewhat using boards of 2 raters doing independent ratings

# Developing Test Items

- Technical Job Knowledge content identified by “What would a new GCOS specifically need to know about, i.e., the programs and practices of the PGC?”
- Discussion resulted in list of specifics using brainstorming, reference materials, and job observations
- Selected the essential and critical
- SMEs worked individually and in groups to draft items

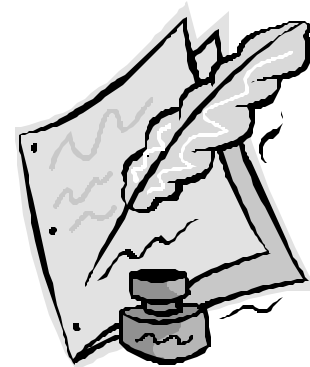
# Situations

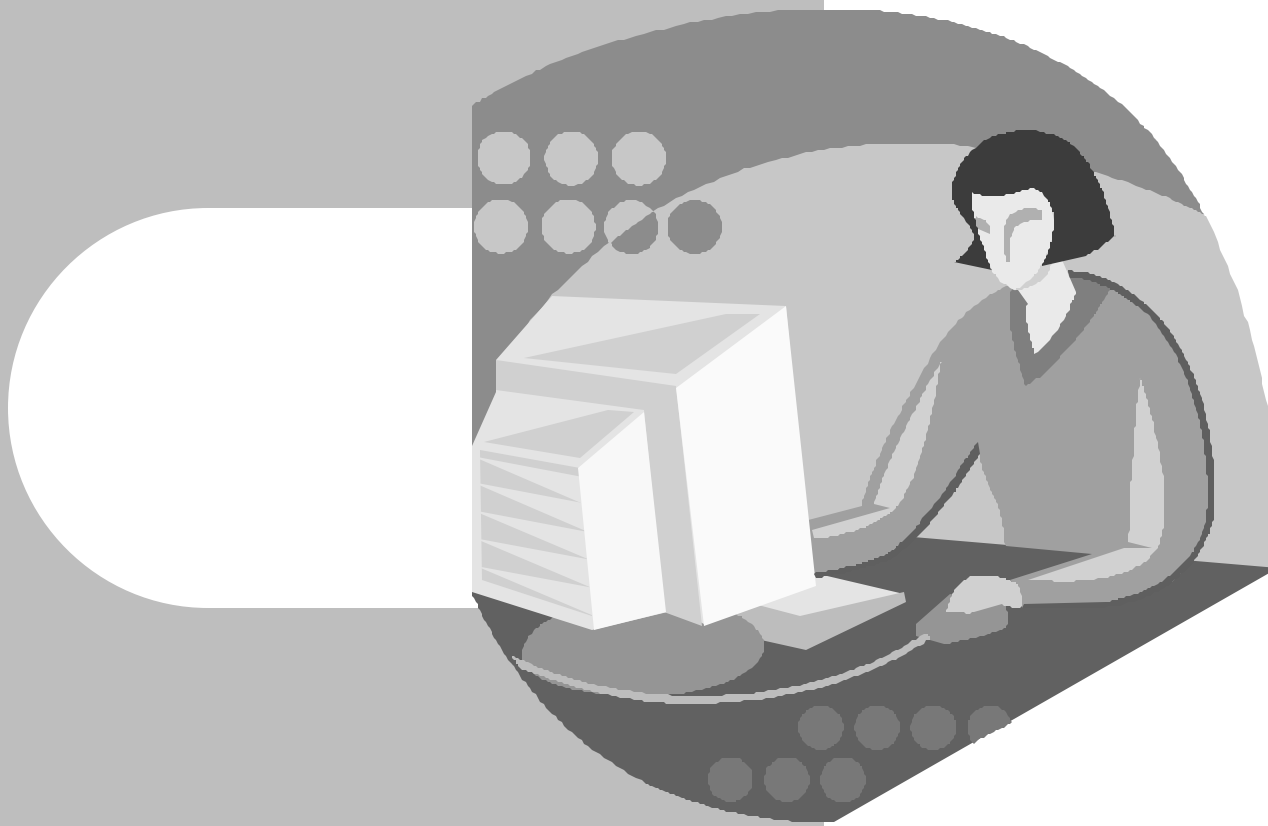


- Effective Working Relationships and Judgment situations identified by “How would a new GCOS apply these abilities?”
- Again, discussion resulted in list of specifics using brainstorming, reference materials, and job observations
- Selected situations that evolved in steps to develop questions

# Written Communication

- Selected situations from both lists and adapted to letter and essay question format
- SMEs wrote responses for rating purposes





## The Test



Now for the good stuff

# Subject Areas Tested

<u>Subject</u>	<u>Questions</u>	<u>Num</u>
Technical Job Knowledge	1 - 35	35
Exercise 1 - The Staff Meeting	36 - 47	12
Exercise 2 - The Home Owners Association	48 - 60	13

# Subject Areas Tested

<u>Subject</u>	<u>Questions</u>	<u>Number</u>
Exercise 3 –Duty Officer – Part 1	61-66	6
Exercise 3 – Part 2 Written Letter	67	1
Exercise 4 – Walk In Complaint	68	1

# Technical Job Knowledge

- **SUBTEST INSTRUCTIONS**

- ✓ 35 questions - measure specific technical job knowledge
- ✓ Select the one best answer choice
- ✓ You may change your answer while on the screen
- ✓ Once you have made your selection you cannot return to the question to change your answer
- ✓ You must answer every question
- ✓ Wrong answers will not be counted against you

# Exercises 1 through 4

The remainder of the test is designed to measure your ability to:

- ✓ Exercise good judgment
- ✓ Solve problems
- ✓ Supervise
- ✓ Relate to others
- ✓ Communicate effectively in writing

# Exercise 1 - Instructions

- Assume the role of a newly appointed supervisor
- You will be given information you may actually encounter on the job
- You must make a decision about what to do with this information

# Exercise 1 - Instructions

- For each question, choose the best response
- When more than one response can be selected, you will see the phrase “Choose one or more” in the question

## Exercise 2 - Instructions

- 3 questions you may encounter as a new supervisor. Score is sum (+) minus the (-)
- 10 questions a supervisor asked by the public. Score is sum (+)
- You must answer every question
- Once you move to the next question you cannot return to change answer

## Exercise 3 - Instructions

- Exercise three is in two parts
- Part 1 = a series of questions based on the developing scenario
- Must make decisions on what you know
- Part 2 = you will write a letter

# Exercise 4

- Essay Question
- Human Raters
- Four Factors



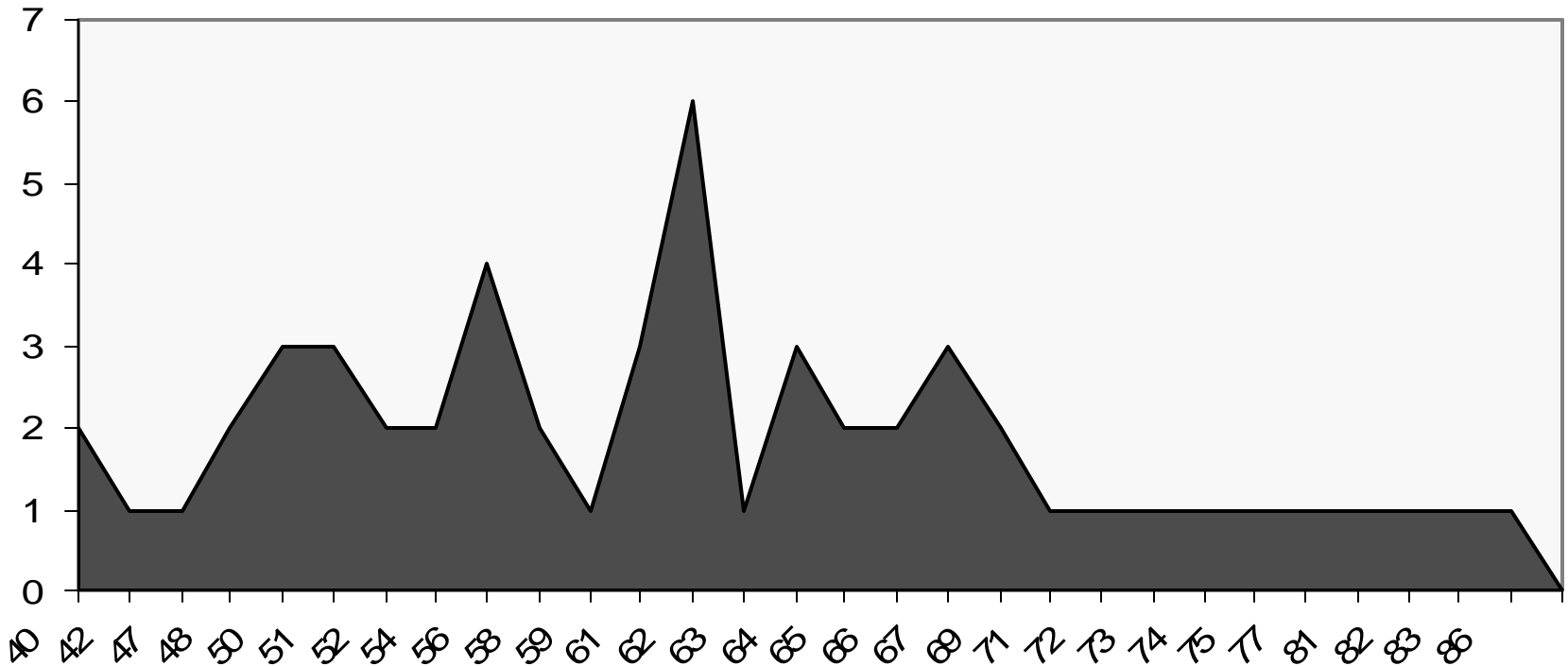
# Scoring



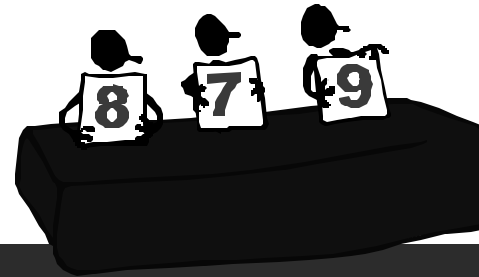
- Technical items: +1 or 0
- Situations: +1 or -1
- Letter: 0 – 8, combined scores of 2 raters
- Essay: 0 – 32, combined scores of 2 raters
- Maximum 110 points
- Candidate scores ranged from 47 to 86
- Mean = 62.72, Median = 62, Mode = 62
- Std. Dev. = 9.2

# Score Frequency Distribution

Frequency Distribution



# Scoring Issues



- No computer-generated item analysis available
- Limited manipulation of data
- Subtest scores copied from candidate result screens, then entered in SPSS
- Passing point determination
- Block-scoring vs. Ranking on list
- Computer interface/human errors

# Results



- Subtest to subtest correlations indicate each subtest independent of others in measurement – This is good!
- Highest correlation between Effective Working Relationship Subtest and Essay Question (.27)
- Lowest between Technical Job Knowledge and Essay (-.08)

# Subtest Correlations

	TJK	STAF	HOA	CITAT	LETR	ESSY
RAW	<b>.29*</b>	<b>.49*</b>	<b>.20</b>	<b>.31**</b>	<b>.34**</b>	<b>.76**</b>
TJK		<b>- .07</b>	<b>- .03</b>	<b>.21</b>	<b>- .09</b>	<b>- .08</b>
EWR			<b>- .02</b>	<b>.01</b>	<b>.27</b>	<b>.10</b>
COM				<b>-.09</b>	<b>.04</b>	<b>.06</b>
JUD					<b>-.10</b>	<b>.00</b>
LET						<b>.14</b>

# Correlations by Scores on Factors

	TJK	EWR	JUDG	WRIT
EWR	- .07			
JUDG	.21	.01		
WRIT	- .08	.10	- .10	
RAW SCORE	.29	.46	.31	.33

# Correlations by Type of Measurement

	Written TJK Subtest	Situational Judgment Subtests
Essay Subtests	.07	.12
Situational Judgment Subtests	.07	

# Reliability Issues

- Good agreement of independent ratings on both written exercises
- SJT and heterogeneous testing methods don't lend themselves to split-half or internal consistency reliability
- Test-retest? Maybe later...

# Validity

---

- KSAs = high content validity
- Items = high face validity and potentially construct validity through test designed as a logical application of KSAs

# The Pitfalls



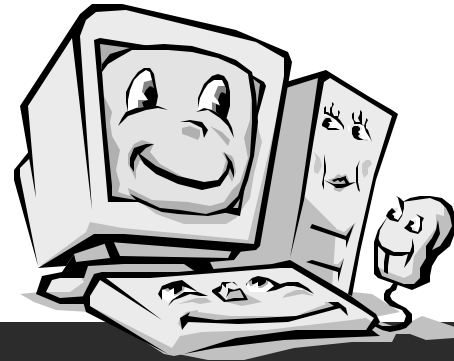
- Labor relations involvement
- Scenario development Trial-and-Error
- Scoring determinations
- Test developed (May – July 2000)  
before we knew how ICE would work  
(April 2002)

# The Pitfalls (cont'd)

- Multiple test administration delays
- Discovered character limitation for responses to human rater questions
- Test administration, scoring and candidate evaluation policies changed during ICE implementation



# The Positives



- Multiple-select/Multiple-choice items with differential scoring made a difference
- Difficulty levels of all subtests comparable
- Test was easy to administer and score
- Real-life job-related situations increased face validity and candidate acceptance
- Future development of computer-administered SJTs will be easier

# Next Time...

- Consider a scoring continuum for all items
- Consider eliminating the human rater questions for ease of administration
- Increase difficulty of Technical Job Knowledge items



# Questions

