File name: N:\OrgsProf\Ac\2024_IPAC\Presentation_Tatum_Issues\Post_Conference_IPAC2024\Wiesen_2024_IPAC_Solutions_to_Tatum_Issues_w_speaking_notes.wpd

1. The slide # is at the top of the page. A box contains the slide, followed by speaking notes.

Possible Solutions to Content Validity Challenges in Court;
An Insider's Analysis and Insights

Joel P. Wiesen, Ph.D.

Contact: jw@jpwphd.com

2024 IPAC Annual Conference
New Orleans, LA, 7/29/2024

Speaking notes:

Good morning. It is a pleasure and an honor to be here.

Great to see you here.

Please be prepared for some new ideas.

Let’s jump right in because there is much to cover.

2.

Who Am I?
Independent consultant
Worked for the MA civil service agency
A scientist-practitioner
46 year member of IPAC
45 years as expert for Defense and Plaintiff
Expert in *Tatum* case for 17 years

I'm an independent consultant.

Out of grad school, I worked for the MA civil service agency for 15+ years

I have worked as an expert for 45 yrs, but not every year. In my role as expert, I have taken a fresh look at some typical testing practices and issues raised in court.

Recently, I served as the main expert for the plaintiffs in *Tatum.*

PACE exam about 1980; Teal v Conn 1982; Expert in a NYC FF cases won on summary judgment (for plaintiffs)

3.

Goals of this Presentation
Identify testing weaknesses seen in Tatum
Also weaknesses seen in other cases
Propose ways to address these weaknesses
Propose new approach to promote police

I'll identify testing issues that I saw in Tatum and related cases, and propose ways to address these weakness. I'll also present some novel ideas to improve test development.

At trial, it was not possible to change what was done, so the main focus at trial was not on solutions.
Now, post decision, we can explore ways to address the weaknesses

Some of these testing issues are little discussed in the literature but can have a huge impact on test validity and validation research.

Finally I'll propose a completely new approach to developing police promotional exams that I think will address many of the weaknesses and improve job performance and perhaps reduce adverse impact.

This was supposed to be a 60 minutes presentation but was scheduled for 45. This will be a fast moving presentation.

I'll try to put my speaking notes on my website. The URL will be on one of the last slides today.

4.

What is the *Tatum* Case?
Minorities challenged 8 sergeant exams
2 exams for Boston
6 annual "statewide exams"
MC & E&E
Used for 100+ PD in Massachusetts
Plaintiffs prevailed
Detailed court decision (75 pages)

The *Tatum* case was a challenge to 8 promotional exams for police sergeant:
2 exams were for the city of Boston
6 exams were so called statewide exams.
Any of 100 PD in MA could choose to participate in an annual statewide exam.

The exams consisted of a MC test and a rating of Education & Experience.

The plaintiffs prevailed in *Tatum*.

5.

Role of a Police Sergeant
Sergeants supervise officers (about 7)
Sergeants spend much time in the field
    Go to the most serious incidents
No sergeant at the Geo. Floyd incident
    Was a minor crime of passing a fake $20 bill
Officers ask sgts. if uncertain (law, SOPs)
Sgts must answer quickly (instantaneously)

A sgt provides supervision to POs in real time by radio and in person.
Sgts go to serious incidents.
There was no Sgt at the Geo Floyd incident since it was a minor incident: passing a fake $20 bill

Sgts need to size up incidents quickly and accurately.

Often Officers ask sgts how to act, and Sgts must answer immediately

Often the questions are legal or procedural.
For example, Sarge, I'm at a domestic dispute. The wife says her husband threatened to kill her. I did not hear him say that, so I don't think I can arrest him. But if I leave him here and he kills his wife, it will look bad for the PD. Can I arrest him? What should I do?

Another example, Sarge, a man and women both say they have sole custody of a young child. What should I do?

Sgts need to know law and PD SOPs, and must think fast.
Sgts also do paperwork and review PO paperwork.

6.

Tatum Decision: Discrimination
"Overwhelmingly persuasive evidence proves that HRD interfered with the class members' rights to consideration for promotion to police sergeant without regard to race or national origin."

The decision is 75 pages and closely reasoned.
I'll quote some highlights, but I encourage you to read the whole decision.

One of several conclusions by the judge:

"Overwhelmingly persuasive evidence proves that HRD [the state Civil Service examination agency] interfered with the class members' rights to consideration for promotion to police sergeant without regard to race or national origin."

7.

Decision: Conclusion

"…a discriminatory system that has injured qualified candidates and **deprived the public of the benefits of having the best-qualified police sergeants**."

The court went on to describe the exam as:

"…a discriminatory system that has injured qualified candidates and deprived the public of the benefits of having the best-qualified police sergeants."

8.

Summary of Tatum Decision

JK tests consistently had adverse impact
Intent discrimination based on past impact
JK tests measured rote memorization
JK tests did not measure important KSAPs
JK tests invalid, especially for ranking
Did not use alternatives with less AI

The court's decision can be summed up by this list of findings:

The state used the same approach for many years and knew it would have severe adverse impact.

The MC test measured mostly rote memorization.

Many important KSAPs were not measured.

Thus, the exams were invalid, especially for ranking.

The state knew about, but did not use, alternatives expected to have less AI.

9.

Decision Example 1
"According to CP, the most critical determinant of future success as a community policing Officer is:
A. Superior communication skills.
B. Empathy.
C. Autonomy.
D. Analytical ability."

The court quoted a number of test questions in the decision. Consider this question:

"According to [the textbook] Community Policing, the most critical determinant of future success as a community policing Officer is"

The key is Empathy.

10.

Decision Example 1
Defense said this item measured empathy
The court critiqued this item as only measuring knowledge about empathy but not the ability to be empathetic or foster empathy in subordinates

The Defense said this item measured empathy.

The court critiqued this item as only measuring knowledge about empathy, but not the ability to be empathetic or to foster empathy in subordinates.

11.

Decision Example 2
"…the exams did not test many important job qualifications."
"…not measure ability to apply knowledge practically and to exercise judgment on that topic in specific situations"

Beyond critiquing individual items, the court was concerned about the KSAPs that were not measured by the exam, writing:

"…the exams did not test many important job qualifications."

"…not measure ability to apply knowledge practically and to exercise judgment on that topic in specific situations"

12.

Items: Court
"Most of the questions on the exams at issue in this case tested topics that were important to the job of sergeant. That does not mean that HRD's format was reasonably job related. It was not."

The judge really did not like some of the items even if the topics seemed relevant.

"Most of the questions on the exams at issue in this case tested topics that were important to the job of sergeant. That does not mean that HRD's format was reasonably job related. It was not."

13.

Related Police Promotion Cases
Two closely related impact cases in MA
Lopez v Lawrence, 2014 (trial in 2009)
Same claim, but in federal court
Court ruled exams were "**minimally valid**"
Smith v Boston, 2015
Same claim in federal court but for lieut. exam
Court ruled 2 Boston exams were invalid

Two other MA police promotion impact discrimination cases are very similar in terms of the:

JA, test outline, item types, and test components.

In Lopez, the same 8 exams were challenged in federal court and the court ruled they were MINIMALLY valid.

In Smith, a different, federal judge held that 2 similar Boston lieutenant exams were invalid.

14.

Smith 2015
"… the Court has found that … too many skills and abilities were missing from the 2008 test outline."

With very similar test development to Tatum, a federal judge ruled 2 Boston Lieut exams invalid.

The court thought the exams covered too little of the job.

15.

Issues and Proposals
Some major issues/flaws in testing
Some little discussed
Based on Tatum and other exams I reviewed
Approach of the rest of this presentation:
Describe issue
Offer a proposed solution

I have had the good fortune to review exams from some of the most respected consultants and largest municipalities and some states.

Tatum and related court cases reveal some major testing issues, some little discussed.

I will describe each issue and offer a proposed solution.

16.

Topics of Issues
Items (7 issues)
Job analysis (6 issues)
Test outline and misc. (7 issues)
New approach to promotional exams (1 issue)

The 20 case-related issues fall in 3 broad areas:

Items
Job analysis
Test outline and misc.

Plus a new approach to promotional exams

17.

1. Issue: Allowing Item Appeals
"…no credit for ... community policing or involvement in the communities served"
"…no credible support for the notion that a bachelor's degree was the equivalent of six years job experience."

The first issue is a very practical one, but one with vast psychometric implications.

Some (or many) civil service systems allow a candidate to appeal items the candidate thinks are faulty.

Candidates like the appeal process because the promotional examinations are high stakes: their only path to promotion. Candidates spend months studying and do not want any faulty items on the test.

Appeals typically are heard by people with no training in testing.

The review body usually upholds an item if it closely reflects the source.

So, over time, item writers tend to quote sources extensively.

As a result, items measure the recall of wording in the sources.

There is little measurement of practical application of a knowledge.

18.

Proposal: Allowing Item Appeals
New law/rule to grant authority to SMEs
e.g.: Post-test agreement of 3 SMEs presumed to be adequate support for an item (SMEs who did not write the item)
Involve police academy, municipal attorney
Concern: Candidates will say they cannot study for SME questions

This issue calls for a procedural solution.

Modify the civil service rules to recognize the role of SMEs in defending item content.

For example, a rule saying post-test agreement of 3 SMEs will be presumed to be adequate support for an item, if those SMEs did not write the item.

Of course, we will need to give clear instructions to the SMEs reviewing the challenged items.

Candidates will not be happy about this. They will say there is no way to study for an exam based on SME opinion. We'll come back to this later.

19.

2. Issue: Replicate Legal Cases
Law questions are common on police exams
Questions typically completely replicate all the facts of an actual court case
This avoids appeals
Does not replicate job duties of sergeant
Officers must respond to incidents that do not fully replicate past court cases

This is a special case of the previous Issue.

Law items are often written based on actual cases and the questions typically replicate all the case facts. This is done to avoid appeals. If not, a candidate could claim that if a court case had different facts, the outcome could be different.

Such items test memory of actual court cases, not application to new situations.

But Sergeants have to make judgements applying legal precedent to new situations.

So the law questions do not involve what a sergeant actually does on the job, despite a superficial appearance of doing so.

This extremely important aspect of the job is little tested.

20.

Proposal: Replicate Legal Cases
Use items that do not fully replicate past cases
Require judgment in applying precedents
Involve local attorney in item development
City attorney, etc.

The proposal is to avoid law items that completely replicate past cases.

Rather, write items that require the test-taker to apply legal precedents to new situations.

Involve legal experts to help write and review the items, again with proper instruction.

Getting enough time from a busy city attorney or district attorney may be a challenge but it is essential.

Explain new approach to candidates before exam.

21.

3. Issue: Definition Items
Easy to write & defend definition questions
Knowing a definition does not mean person can use the concept

Definition items are easy to write, and easy to support if challenged.

But often they measure the wrong thing.

Knowing a definition does not mean a person can apply the concept.

It is essential to measure understanding and application, not just memory.

22.

Proposal: Definition Items
Use only a small proportion of definition items for any given KSAP
Use definition item only if the definition is important to know in order to do the job

The proposal for definition items is twofold:

Use only a small proportion of definition items overall and for any given KSAP.

Use a definition item only if implementing the knowledge is straightforward, or if the term is used on the job.

23.

4. Issue: Items on Procedures
Easy to ask procedural step order or names
Knowing the correct order of steps or names does not mean person can execute the steps

Q: In the SARA problem-solving model, what should be done in the analysis step?
key: Collect information from a variety of public and private sources
Does not test if able to collect information

Similarly, it is easy to write items concerning the name or order of steps in a procedure.

But knowing the names of order of the steps does not indicate a person can execute the steps properly.

For example, in problem solving, knowing the first step is collecting information from a variety of public and private sources does not mean the test taker can actually collect information.

SARA model from SWANSON_TERRITO_TAYLOR_7th_ed_2008_POLICE ADMINISTRATION.txt
Scan, analyze, respond, assess

24.

Proposal: Items on Procedures
Test ability to implement the steps
Test the name or sequential order of steps only if these are important (e.g., step is likely to be done out of sequence)
Use only a small proportion of step name or sequence items for any given KSAP

Items on procedures should test the ability to implement the steps of the procedure, unless the names or sequence of the steps are as important as implementing the steps.

25.

5. Issue: Academic Items
Question on desired leadership style
Key: balances concern for people and task
A correct answer does not mean the person can do either well

Often college textbooks are on the list of required readings for police promotional examinations. College textbooks typically have academic treatments of topics (with full histories of many topics, often going back 100 years). This has pros and cons.

A question on leadership style might have as the key that the leader balances concern for people and task.

Answering that question correctly does not mean a person can do either well.

In general, academic items are a large step removed from performing the job.

Often the long history of a topic is of no help when doing the job.

26.

Proposal: Academic Items
Use academic items only if application clear
Items on important, applied topics
More situational questions
    Video stimuli
    Constructed responses
May require item writers to have more skills
    Work with actors and video content creators

I propose using academic items which have a clear, straightforward application and writing *situational* questions that test the information as applied on the job.

Situational questions might require video stimuli and constructed responses.

Such items might well require more time and expense and even working with actors and video content providers.

We may need to find or develop source material that support such questions.
- The critical incident approach could be of help in developing such material
- Will return to this topic at the end of this presentation.

27.

6. Issue: Limited Item Review
SME item review questions, 2008 exam:
Suitability for rank?
No definition of suitability or suitable
Estimated difficulty "for the persons taking the examination"
Estimated readability

Too often the SME review of test questions is limited in scope and without adequate structure as to what is rated and the rating choices.

In Tatum, SMEs who reviewed items as part of the test development process were asked:
1) Is the item suitable for the rank?
No definition of suitable
Various SMEs can have very different views of suitability

2) Estimated difficulty rated but "for the persons taking the examination."

The level choices were
- easy "71 to 100% of applicants will likely respond correctly"
- medium
- hard "0 to 40% of applicants will likely respond correctly"

No link to job, just to the estimated applicant group

3) Estimated readability
But SMEs have no credentials or training to answer this question.

28.

Proposal: Limited Item Review
Gather better information from the SMEs
Improve the item rating form
Clarify the review topics and rating levels
    Is this K important to do the job?
    Is this the best way to measure this knowledge?
Talk with SMEs about each item.
    How is this knowledge used on the job?

I propose to gather better information from SMEs
Be more precise in the task given to item review SMEs.

Clarify the review topics, carefully word the questions and rating choices. For example, ask SMEs:
Is this Knowledge important to do the job?
Is this the best way to measure this Knowledge?

Task or KSAP questionnaires now do not capture enough information to support test item writing.

Supplement the JA with discussion with the SMEs about each item.
How is this K used on the job?

29.

7. Issue: Job Analysis Accuracy
Task and KSAP inventories with implausible results
Tasks not done daily; KSAPs omitted
Major disagreement among SMEs
Illogical ratings:
Tasks of budgeting; read, interpret tables/graphs: but no math ability required

JA accuracy can be low.
It is tempting to simply base the test outline on the mean ratings in a JA report. But sometimes there are glaring indications that the ratings may not be trustworthy:
- Tasks are said to be done daily but are clearly not daily
- KSAPs that are of obvious importance are omitted or rated as not needed
- Extreme differences in ratings between SMEs

I have seen illogical ratings. E.g., one Management job included Budgeting, and read & interpret tables & graphs. But SMEs said no math ability was required. You can't read & interpret tables and graphs without math ability.

As another example, FF SMEs who rated Fleishman areas testified in court that they were confused by, and just did not understand, the Fleishman areas when they did the ratings.

In the section "Analysis of Work" of the SIOP Principles (the term now used for job analysis), "Lack of consensus about the information contained in the analysis of work should be noted and considered further." (Page 7, col 2, last par, 2018).

30.

Proposal: Job Analysis Accuracy
Do not blindly rely on SME ratings
Probe discrepant and suspect ratings
Review the JA results for plausibility
Conduct reviews of JA findings with SMEs
Gather ratings on KSAPs from
- Supervisors
- Training academy staff

One proposal is straightforward: Do not blindly rely on SME ratings.

Conduct reviews of JA findings with SMEs and track down the reasons for any large differences in ratings.

Also, job SMEs may not be KSAP experts. I propose going beyond using incumbents to rate KSAPs. Also collect KSAP ratings from Supervisors and Training Academy staff.

I think that incumbents are used as SMEs because long ago one court once said that incumbents know the work better than supervisors. I think that has led our profession down an unfortunate path, since incumbents do not think much about KSAPs.

31.

8. Issue: KSAPs Not Well Defined
(A) K of principles of management
versus
(B) POSDCORB areas listed separately
If (A), are all SMEs rating the same area?
If (A), how much emphasis on each facet?
Lack of clarity affects job analysis
Lack of clarity affects test outline

The KSAPs listed on JA inventories are typically short phrases.

A and B on this slide show two ways to present a K.
W approach A only 1 K statement is on the JA questionnaire - namely K of Principles of Management

Recall: POSDCORB stands for:
Planning, organizing, staffing, directing, coordinating, Reporting, Budgeting

With approach B, 7 Ks appear on the JA questionnaire.
If approach A is used, various SMEs might focus on any one or several of the 7 Ks. So, the SME ratings might make little sense as the SMEs have idiosyncratic conceptions of what it is they are being asked to rate.

Approach B might lead to more exam weight being given to management than budgeting, esp for the job of Sergeant, reflecting the job. Approach A may lead to incorrect emphasis on budgeting, especially for lower ranks.

32.

Issue: KSAPs Not Well Defined
Consider these 3 K statements from a recent job analysis for police lieut. and captain:
Principles of police administration
Supervision, management, and leadership principles
Community-policing and problem-solving principles
What do these Ks cover, what is being rated

What does the first K cover or include? Everything in a 600 page text on police administration? Or a small subset of that

Are all community policing practices equally important? What these K statements cover? Note, there is no rating of specific methods, techniques, or practices. Notice that the word "practices" is not there.

That means we are not sure what should be tested.

When the same K statements appear on the JA questionnaire for Lieut & Captain, is the K area the same for both ranks?

If the JA lacks specificity, asking SMEs to interpret the JA results after the fact may not work well. That would be relying on one or a handful of SMEs rather than on the larger N that completed the questionnaire.

33.

Proposal: KSAPs Not Well Defined
Ask if every SME will agree on KSAP scope
Ask if the KSAP can be broken down into components that are not highly correlated
Try using operational definitions of KSAPs

I propose the following to better define the KSAPs listed on a JA questionnaire and a test outline.

When developing the questionnaire, ask whether all SMEs will have the same conception of the KSAP scope. If not, break it down. If not sure, clarify the description.

Also, if the KSAP can be broken down into components that are not highly correlated, it should be broken down.

Consider the approach recommended in the Uniform Guidelines: operational definitions of KSAPs

34.

9. Issue: Many Tasks/KSAPs
Often there are many tasks and KSAPs
Group tasks into categories loses detail
Grouping KSAPs into broad competencies loses detail

Another job analysis issue is that often there are many tasks and KSAPs.

That makes both the JA questionnaire and the test outline unwieldy.

One approach that is sometimes used is to group tasks into a smaller number of duty categories and KSAPs into larger competencies when developing the test outline.

However, such grouping loses detail. If you want to link your test items to the job but you're only linking the items to duty categories or competency areas, you've lost valuable detail of the tasks and the KSAPs.

35.

Proposal: Many Tasks/KSAPs
Use tasks and KSAPs when writing items
Do not rely on task/KSAP groupings

With respect to the JA questionnaire, give SMEs a break now and then to keep them fresh. Or divide up the JA questionnaire over days. Include checks for random responding. E.g., include nonsense tasks or KSAPs. E.g., repeat some tasks/KSAPs and see if ratings are consistent. We need good JA data!

With respect to the test outline, the proposal is to link test items to individual tasks and/or KSAPs. Don't rely on task and KSAP groupings when trying to prove content validity. Groupings are to general to guide test development.

36.

10. Issue: KSAP Weight
Easy to write & defend definition questions
Knowing a definition does not mean person can use the concept

Most JA questionnaires use ordinal rating scales, such as a five-point scale ranging from 1 = not important at all, to 5 = important to a very large extent

Calculating means of ordinal data violates assumptions.

Beyond that, there is no standard scope or size of a KSAP.
Recall:
Principles of Management versus the 7 topics POSDCORB

Arbitrary decisions on KSAP wording can greatly affect the distribution of items on a test outline.

Also, some sources lend themselves to item writing, and other sources do not.

Further, there are comprehensive sources for certain KSAPs and not for others.

All this makes developing a test outline a messy, inexact process.

37.

Proposal: KSAP Weight
Rate KSAPs with **ratio** scale
How much of successful job performance depends on this KSAP?
Allot 100 or 1,000 points among the KSAPs
Use Excel to ease math burden
Frank Landy used this approach in his job analysis of police officer in Massachusetts

I suggest the SMEs rate KSAPs and required reading material using a ratio scale.

A direct question can be asked, such as:
How much of successful job performance depends on this KSAP?
How much of job performance depends on this task?

Frank Landy used a ratio scale for a job analysis he did for the job of Police Officer in MA many years ago. He had the SMEs divide up 100 points among the 20 Fleishman areas.

Alternatively, use a ratio scale such as:
1 - lowest importance
2 - twice as important as 1
3 - three times as important as 1
4 - etc.
Then you can calculate the proportion based on sum of all ratings.

38.

11. Issue: Past Job Performance
Calls to measure past job performance
    Fields (2007) PTC Presidential Message
Empirically keyed biodata has high validity
    In top 3
    Sackett, Zhang, Berry, and Lievens (2022)

Cassie Fields 2007 PTC Pres message:
Police Mgt wants to give credit for past job perf

A Police Chiefs has told me the same.
- "My two best POs cannot get promoted."

Sackett. Zhang, Berry & Lievens (2022), Table 3:
Employment interviews, structured 0.42
Job knowledge tests 0.40
Empirically keyed biodata 0.38

39.

11. Proposal: Past Job Performance

Body Cam Review
- By outside raters
- Videos provided by candidate and supervisor

Accomplishment Record
- Fields (2007) PTC Presidential Message

Job Performance Evaluation
- Many articles
- Landy (1977) Police Foundation Report

The main resistance to including past job performance on CS exams has been takers' concerns of bias. We now have new technology. We should explore using Body and Cruiser Cams

- contributed by taker

- contributed by supervisor

- Perhaps of specific types of interaction

40.

12. Issue: Critical Incident Usage
List of tasks (briefly stated) is inadequate
Short "ride-a-longs" are inadequate
Often critical incidents are not collected

It is basically impossible to understand a job and write job-related items based on a list of important tasks and KSAPs.

The short ride-alongs often done are a good start but are too brief, and therefore inadequate.

Unfortunately, collecting critical incidents seems to have gone out of style.

Often few or no critical incidents are collected.

Doverspike and Arthur said just this in 2012 book chapter on "*The Role of Job Analysis in Test Selection and Development*": "Test development cannot be based ... on task information or KSAO information alone. To adequately document test development efforts, it is necessary to collect and document detailed information on tasks and KSAOs."

41.

Proposal: Critical Incident Usage
Collect critical incidents from incumbents
Collect critical incidents from supervisors
Goal: Many hundreds of incidents
More nuanced understanding of the job
Incidents provide grist for item writing

Critical incidents can provide rich information about the job duties and qualifications. These are easy to collect.

For example, I collected a few hundred for FF by just having the FD send out a notice with a few attachments.

Critical incidents can put meat on the skeletal lists of important tasks and KSAPs provided by JA questionnaires.

Critical incidents also can serve as springboards for developing simulations.

42.

13. Issue: Outline Based on Tasks
Typically test outlines are based on KSAPs
KSAPs are a step removed from the job
Sources for KSAPs can be quite academic
How to tell if a test is representative of job?
- Covering KSAPs ≠ covering tasks
- KSAP importance may not map to task criticality

Typically test outlines are based on KSAPs, but KSAPs are often inadequately defined.

Sources related to KSAPs can be quite academic (esp. textbooks) and thus of questionable linkage to job tasks.

If you are trying to prove to a court that a test is representative of a job, links to job tasks are very convincing. Links to KSAPs less so since KSAPs are a step removed from the job.

Incumbents actually do job tasks.

The KSAPs required to do the job tasks is a step removed from the job and can be a matter of opinion.

43.

Proposal: Outline Based on Tasks
Develop test outline based on tasks
Can have 2-way outline
sources and tasks
KSAPs (or KSAP groupings) and tasks
Easier to show test is representative of job
Ask SMEs/candidates about missing topics

The proposal for test outlines is simple: include tasks in the test outline.

Perhaps have a two way test outline of sources and tasks
or
KSAPs and tasks

A 3-way outline that includes tasks, KSAPs, and sources might be ideal but that seems cumbersome.

I think including individual tasks in the test outline will help support a claim that the test is representative of the job, more so than just a KSAP-based job linkage.

44.

14. Issue: Weighting by # Items
Important topics get more questions
Score on exam is typically # correct
Easy and hard items have same weight
Some topics tested with few items
No **reliable** measure of such topics

We use a simple-minded approach to scoring tests: the score is simply the number correct.

Easy and hard items have the same weight, as do items on different topics, some topics more important than others.

Further, some topics are tested with as few as 1 item

There is precious little reliability measuring a K with one or two items.

In one exam with 100+ test takers, test reliability was ranged from .04 to .58 for subtests with 4 to 10 items

45.

Proposal: Weighting by # Items

Validity capped by square root of reliability
Valid topic w/ few items ➡ invalid measure
Enough items to reliably measure a topic
Minimum of 10 items per area
Weight topic scores by importance

I propose we strive for a reliable measure of each test topic that appears on the test outline.

That requires having a minimum number of questions per test topic.

We can weight each test topic in a supportable fashion to calculate a total score.

This will require longer tests and more testing time. Two or more days of testing might be needed.

Better longer tests than using unreliable measures.

(Spearman Brown prophecy formula explorations led to my rec of min of 10 items.)

Source: *Spearman.Brown.Prophecy.Formula.xls* in statistics folder

15. Issue: Item Weight
We now weight all items equally.

Equal weighting of items is easy to do, but it is illogical.

Some questions ask about really central topics and other questions are more peripheral. Why should such different items get equal weight?

That we have always done it that way is not a good reason.

If we do not improve testing methodology, we will never raise the low levels of validity that we have long seen as a low ceiling.

47.

15. Proposal: Item Weight
Weight items by consequence of error.

It would seem to be easy to ask SMEs to rate the consequence of not knowing the correct answer to an item.

It is worth a try.

48.

16. Issue: Setting Passing Point

Angoff rating is compensatory

An essential area might have all easy questions

Illusive "minimally qualified incumbent"

This is indirect rating of passing point

Assumes that because a minimally qualified person knows the topic, the topic is required to do the job!

The Angoff method is widely used but has logical weaknesses and can yield unusable results, with no passers. When Angoff fails all candidates, some consultants set the passing point at 2 s.d. below the Angoff point, with no justification.

It seems the Angoff process seeks to pass a replica of the current minimally acceptable workforce. But it does so in an ineffective fashion. If 80% of incumbents know X and 40% know Y, it might be that the new passers have those levels of knowledge reversed.

Also, Angoff assumes that one passing point is reasonable for all tested topics. If few incumbents know topic A, that will lead to a low pass score for topics A & B, even tho most incumbents know topic B.

The logic of Angoff approach is indirect. It assumes that because a minimally qualified person knows the topic, the topic is required to do the job.

Angoff also assumes all items have equal worth, but clearly that is not true. Some items measure more important KSA's.

49.

Proposal: Setting Passing Point
Consider alternatives to Angoff procedure
Rate exam as an entity
How many items answered correctly indicate a person can do the job?
Do this by test area
Consider contribution to job performance

It is time to explore alternatives to the Angoff.

We could rate the exam as an entity rather than rate individual test questions. The total is not necessarily a simple sum of the items.

We could ask, "How many of these items would have to be answered correctly to indicate a test taker can do the job."

Alternatively, we could rate the items on each test topic, and have a passing point for each tested topic. More on this shortly.

We could think about additions to the Angoff process, such as ways to add the amount the item contributes to job performance or the consequence of not knowing a fact.

50.

17. Issue: Single Pass Point
Items weighted equally
Grading compensatory
Can pass exam with zero on a KSAP
If only one person passes exam ➡ promote

The previous few issues raise the issue of using a single passing point.

A single passing point assumes all test questions contribute equally to validity and avoidance of serious errors on the job. That means all test topics are of equal import and easy and hard items provide equal information.

With a single passing point, it is possible to pass with zero knowledge of an essential KSAP.

The passing point is important because, in some munis, only one person passes, often with a low score, and that person will be promoted.

51.

Proposal: Single Pass Point
Identify essential KSAPs
Set passing point for each essential KSAP

The proposal is simple: Require a passing score for each essential KSAP, as multiple hurdles.

The danger is that no one passes. But the larger danger is that unprepared, incompetent candidates are promoted.

We'll come back to this at the end of this presentation.

18. Issue: Test Outline Secrecy
Often testing groups do not reveal outlines
Often sources are voluminous
Candidates do not know what to study most
Identifying material to study not job related

The sources on police promotional exams are often voluminous.

Many textbooks contain much information that never appears on promotional exams.

If the test outline is not divulged, candidates have to guess where to put their study time. That guessing ability is not a required KSAP for job performance.

53.

Proposal: Test Outline Secrecy
Tell candidates the # items per source
    Allows candidates to apportion study time
Give guidance on what will not be tested
    History older than 10 years
    Chapters x, y, z

I propose we tell candidates about the test outline: Especially the number of items or weight per source or test topic.

Alternatively, tell the candidates what sections or types of information will not be tested, with many examples.

54.

19. Issue: Professional Secrecy
There are no compendia of:
- Test outlines
- BARS scales
- Practical exercises

Secrecy concerning test development methods and tests is a serious roadblock to advancing our profession.

There are no published compendia of
Test outlines
BARS scales
Practical exercises

And few published compendia of MC Job Knowledge Test Questions

Some of these are available in the field of education, but not for employee selection.

That means each test developer has relatively limited resources to draw on when embarking on a test development project for employee selection.

55.

Proposal: Professional Secrecy
Perhaps IPAC could publish compendia

Perhaps IPAC could help address this deficiency by publishing such compendia.

Perhaps discuss this at the business meeting.

56.

20. Issue: BARS Reliability
Within board reliability: 0.97, 0.91
Between board reliability: 0.44
But .97 x .91 = .88 (n>100)
Within board reliability: 0.97, 0.94
Between board reliability: 0.66
But .97 x .94 = .91 (n>100)
(Note: data not from MA)

This issue concerns the reliability of BARS scoring of oral exercises.

Rater reliability within a rating board is often high.

Despite within board reliability of well into the point 90's, the between board reliabilities were .44 and .66.

Between board reli should have been > .80 since .9 * .9 = .81.

Also, sometimes see high raw score discrepancies between pairs of raters. For example, one rater gives the highest possible score on oral communication and other rater gives an unsatisfactory score (raters from two different rating panels) on an ability such as Oral Communication.

Proposal: BARS Reliability
Use duplicate rating boards
Research into reasons for disagreement
Within a board
Across boards

I propose we use multiple rating boards more often, even routinely, when using BARS scales. At least two boards.

We need applied research into the reason for this puzzling difference between within and between board rater reliability.

Possible research:
1. Have boards discuss ratings with largest discrepancies
2. Have monitors switch panels to look for reasons for diffs
3. Look at subject matter of questions with largest and smallest differences between boards
4. Have panels report on answer key refinements post training
5. Compare absolute differences in scores with correlations
6. Your thoughts welcome!

New Approach to Exams
For police promotional exams
Several goals of the new approach
Better job performance of sergeants
Reduced adverse impact
Better acceptance of promotion process
    Candidates
    Management

In this last topic of the presentation, I propose a new approach to developing police promotional exams

The goals of this new approach are:
- Better job performance of sergeants
- Reduced adverse impact
and
- Better acceptance of the promotion process by Candidates and Management

59.

Issue: New Approach Needed
Candidates prepare themselves for promo.
Hard to learn supervision, management, leadership, strategy, tactics, etc. from books
Exam grades show candidates lack KSAPs
Many high paying occupations have training programs
    $200,000+ average gross pay, Boston Sergeant

Reasons why we need a new approach to selecting people to promote in police departments.

First, it is hard to learn supervision, management, strategy, tactics, leadership, interpersonal relations, etc. from books.

Second, exam scores show most or all candidates are ill-prepared for promotion, even on hard knowledges. The exam grades are not in the high 90's

Third, many high-paying occupations have formal training programs; Boston sgt gross pay is $200,000/yr on average, including overtime

Also, if learn based on coaching and job assignments, it may be that "old boy" network results in unfair advantage for some candidates

Missing the forest for the trees. We do what we have always done or what we are asked to do (CS rules, tradition). Ignoring fundamental problem: Current system not working Need to be I/O psychologists not psychometricians

60.

Issue: New Approach Needed
Currently, many/most exam scores are low
A low-scoring candidate may be promoted
Candidates weak in essential KSAPs
No training for newly promoted sergeants

Our current promotion system is not working well

Many, sometimes all candidates get low grades.
- a low scoring candidate may be promoted

Candidates weak in essential KSAPs

No training for newly promoted sergeants
- Candidates prove they are weak in various topics but are not trained before starting new job

The approach I will describe next is similar to one mentioned in an article in the 2006 PTC-MW Quarterly Newsletter. The author was from the City of Macon (w/ 15 year tenure) but I do not know the author's name. (I am missing the first page of the article.)

61.

Proposal: New Approach Needed

Establish a thorough training program for promotional candidates
- Strategy and tactics for incidents
- Planning and resource allocation
- Interpersonal aspects of policing

This is a major undertaking
- Fund course development
- Fund training time for candidates

I propose we conceptualize the police promotion process as IO psychologists rather than psychometricians.

Establish a thorough training program for promotional candidates that covers many topics, such as:

- **Strategy and tactics**
- **Planning and resource allocation**
- **Interpersonal aspects of supervision, interacting with constituents and other groups and individuals**

The curriculum material for these courses could be used as source material on the promotional exams.

Note that very few PDs have any training programs for Sgt, not for candidates and not for newly promoted Sgts.

This is a major undertaking and various obstacles must be addressed, such as: Funding course development
Funding training time for candidates

We will not improve the job performance of police departments unless we change the way candidates for promotion are trained and tested.

62.

In Closing
We can improve police testing practices
Better content valid knowledge tests
Both candidates and management benefit
Avoid or win more testing court cases
Improve job performance of sergeants

I see the legal system as having the potential for improving our professional practice.

We can improve police testing practices:
Measure knowledges in a content valid fashion
Measure many KSAPS that are now little tested.

Candidates, management, and the public would all benefit

We would avoid, or win, more testing court cases

We would improve the job performance of sgts, all superior officers, and PDs overall

63.

Contact Information
Session URL: http://jpwphd.com/ipac2024
Related 2024 SIOP Master Tutorial
Email: jw@jpwphd.com
Telephone: (617) 244-8859 (land/no text)
Email and telephone calls welcome!
Q&A's

Copies of these slides are on the web at the URL shown.

My contact information is shown.
I welcome communication from one and all on any of the topics of this presentation

The floor is open now for Q&A's.

64.

Main Cases Cited

Lopez v. Lawrence
Lopez v. City of Lawrence, Civil Action No. 07-11693-GAO (D. Mass. Sep. 5, 2014)
https://casetext.com/case/lopez-v-city-of-lawrence-1?q=07-11693-GAO&sort=relevance&p=1&type=case

Smith v. Boston
Smith v. City of Boston, 144 F. Supp. 3d 177 (D. Mass. 2015)
https://casetext.com/case/smith-v-city-of-bos-1?q=12-10291-WGY&sort=relevance&p=1&type=case

Tatum v. Massachusetts
Tatum v. Mass., C.A. 0984CV00576 (Sup. Court 2022)
https://www.mass.gov/doc/tatum-et-al-v-human-resources-division-related-superior-court-decision-102722/download

65.

Presentation Citation

Wiesen, J.P. (2024). Possible Solutions to Content Validity Challenges in Court; An Insider's Analysis and Insights. *International Personnel Assessment Council*, New Orleans, LA, United States.