Content Validity Court Case (Tatum, 2022): An Insider's View and Analysis

Joel P. Wiesen, Ph.D.

jpwiesen@gmail.com, (617) 244-8859

2023 Annual IPAC Conference, 7/25/2023

Last Week: Galapagos Islands

Presentation Links

- Tatum decision (posted)
- PowerPoints (yet to be posted)
- Audio recording (yet to be posted)
- **http://jpwphd.com/ipac2023**

Your Questions

- Short, clarifying questions during the talk
- Other questions at the end

Topics of This Presentation

- Decision in Tatum 2022
  - Background
  - Major issues addressed at trial
  - Tatum 2023 (damages- will not cover)
  - Takeaways (direct and indirect)
- Comparison with two closely related cases
  - Lopez v City of Lawrence (2014, heard 2009)
  - Smith v Boston (2015)

Learning Objective 1

- List and describe the major findings of the court decision in Tatum (2022)

Learning Objective 2

- List and describe the positions of the Plaintiffs and the Defense concerning test validity in Tatum (2022)

Learning Objective 3

- List and describe the positions of the Plaintiffs and the Defense concerning adverse impact in Tatum (2022)

Broad Overview of Tatum

- Challenged promotional exams for Sergeant
- Exams spanned 8 years
- One large and many smaller munis in MA
- Civil service exams developed in-house
- Exams consisted of M/C questions and Education and Experience rating (E&E)

Blockbuster Decision

- Long (75 pages)
- Several fundamental psychometric topics
- Intricacies of adverse impact (AI)
- Reliance on SMEs
- Limitations of MC job knowledge items
- Ranking v banding
- Strongly conclusions by seasoned judge

From Decision Conclusion

- "**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."
  [emphasis added]

From Body of Decision

- "…regularly administered written exams, **knowing** that its testing format had an **unnecessary, plain and obvious adverse impact** upon Blacks and Hispanics".
  [emphasis added]

From Body of Decision

- "HRD **knew of clearly superior assessment methods,** but **continued to use the same, unnecessarily discriminatory format** anyway."
  [emphasis added]

From Decision Conclusion

- "HRD **failed to implement some very simple ways to reduce adverse impact** upon Black and Hispanic candidates."
  [emphasis added]

From Decision Conclusion

- "Instead of improving its assessment format, HRD promulgated lists to provide **a thin veneer of apparent justification** for a discriminatory process."
  [emphasis added]

From Decision Conclusion

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

From Decision Conclusion

- "In all these actions, **HRD knew what it was doing."**
  [emphasis added]

What Led to This Conclusion?

- "The court finds … by a **preponderance of the evidence** it finds **credible**"
  [emphasis added]
- This is the subject of the rest of this talk.

Summary of Decision in Tatum

- Used job knowledge tests for many years
- JK tests consistently had adverse 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
- Intentional discrimination based on impact

Decision in Tatum

- 75 page decision
- Decision is on website: **http://jpwphd.com/ipac2023**

Tatum General Background

- Police Sergeant promotional exams
- Used by many (100+) munis, large & small
- Developed by the state agency (HRD)
- Two components: M/C 80%, E&E 20%
- Time in grade as minimum qualification
- Class action: Black and Hispanic takers

Tatum General Background

- Exams held annually
- Munis participate biannually, typically
- M/C questions based on written sources
  - Textbooks for statewide exams, including law
  - Textbooks and SOPs for Boston exams, w law
  - Many **verbatim** quotes from sources
  - Many **definition** questions

Tatum General Background

- Passing score usually 70%
- Difficulty varied across years
- Passing points not correlated with difficulty
- Passing points set, in part, to reduce AI

Tatum General Background

- Education & Experience used point method
- Score of 70% if meet minimum quals.
- All meet min. quals. so all score at least 70

Details of Decision in Tatum

- Topics discussed in decision
- Other topics discussed at trial
- Combining data across exams and munis
- Aspects of adverse impact AI (means, p/f)
- Validity of Education & Experience (E&E)
- Validity of Job Knowledge (JK) test

Pattern of This Presentation

- Defense's position
- Plaintiffs' position
- Court decision

Aggregation: Defense

- Omit PDs with no promotions from statistical analyses of both AI and means.
  - If omit PD for promotion, also omit for means
- Do not aggregate across years because retakers violate the assumption of independence.

Aggregation: Plaintiffs

- Include PDs w no promotions in statistical analyses of means (but not promotions)
- Defense should provide candidate IDs to enable an analysis omitting retakers
  - Candidate IDs never produced in usable form

Aggregation: Court

- "The court rejects [Defense's] position that it should disregard entirely any results that are not based upon a single test in a single department, after excluding all departments that had no diversity or made no promotions. [That] approach is **biased in favor of finding no difference** in treatment of White, Black and Hispanic candidates.

Adverse Impact **Means**: Defense

- No statistically significant difference between mean scores for some years/PDs
- Should omit PDs that made no promotions when comparing Minority and White means
- Should require significant 3 group ANOVA (B, H, W) before testing Minority-White
- Should **not combine** across PDs or years
  - Due to retakers and independence assumption

Adverse Impact Means: Defense

- Illogical footnote to Defense AI tables: "Pairwise (i.e., B-W, H-W and B/H-W) significance of average JKT and EE scores was tested only when ANOVA across all race/ethnic subgroups indicated the existence of statistically significant differences among the race/ethnic subgroups ($p < 0.05$)."

Adverse Impact Means: Plaintiffs

- Mean score differences germane to AI
- The test statistics show a pattern
  - Scores drive P/F and promotion rates
- Power greater for means than promotions
- Many statistically significant differences between M-W means
- Collapsing across munis logical for means
  - Collapse across years thwarted by no cand. IDs

Adverse Impact Means: Court

- "The principal difference is that Dr. Wiesen included departments that did not make promotions. The court accepts ... Dr. Wiesen's methodology which the court finds persuasive."

Adverse Impact **Ratio**: Defense

- No adverse impact for most of the 8 exams
  - No statistical significance in small munis
  - "Shift of one" avoids 80% for small munis
- Collapsing across munis illogical
- Collapsing across years statistically wrong
  - Retakers violate statistical assumption
- **Simpson's Paradox** could be operating

Simpson's Paradox Interpretation

- Simpson's Paradox theoretically possible
  - Potential threat of misinterpretation

Simpson's Paradox: Within Muni

Muni	Group	Outcome		
		Not Promo	Promoted	Promo Rate
Muni A	Minority	1	1	0.5
	White	10	10	0.5
Muni B	Minority	18	2	0.10
	White	45	5	0.10

Simpson's Paradox: Across Muni

Group	Outcome		Total	Promo Rate
	Not Promo	Promoted		
Minority	19	3	22	0.14
White	55	15	70	0.21
AI Ratio				**0.64**

Adverse Impact Ratio: Defense

- Bonferroni used for tests of individual PDs
  - Shows no AI

Adverse Impact Ratio: Plaintiffs

- **Simpson's Paradox not seen in case data**
  - Pattern of ratios the same overall and in PDs
  - No actual misinterpretation of ratios
  - Defense raised issue of Simpson's Paradox to try to mislead court
- Pattern of severe AI for many decades

Adverse Impact Ratio: Plaintiffs

- Statistically significant AI in largest muni
- Wrong statistical test used for small munis
  - Bonferroni illogical for individual PDs
- Low power for munis with only 1 M taker
- Look at pattern of data
  - Means drive P/F and promotion ratios

Adverse Impact Ratio: Plaintiffs

- Use "shift of one" for every muni wrong
- Collapsing across munis logical for promos
  - Mantel-Haenszel test appropriate
- Collapse across years was thwarted by Defense by not providing candidate IDs

Adverse Impact Ratio: Court

- Court decision
  "P-values greater than .05 but well below 1.0 … does not mean that the data lack all meaning, or that a court should exclude the data from consideration as part of a larger body of evidence."

Adverse Impact: Court

- "The massive amount of evidence proving the known and unjustified disparate impact of HRD's format leaves no doubt in this court's mind…"

Adverse Impact Ratio: Court

- Court decision
  "plain and obvious adverse impact upon Blacks and Hispanics"

Education&Experience: Defense

- Point-year system of E&E
- E&E claimed to measure many KSAPs:
  perceiving & reacting to the needs of others
  ability to write, prepare reports
  ability to be confidential
  ability to follow policies and procedures
  ability to interpret policy

Education&Experience: Plaintiffs

- Years of exp not tap specific KSAPs/duties
- E&E not linked to specific KSAPs or duties
- Amount not quality of experience credited
- No credit for experience outside a PD
- Max impact of E&E on grade is small:
  Mathematically: 6 points out of 100
  Practically: a 9 points due to M/C guessing

Education&Experience: Plaintiffs

- Not content valid
- No valid basis for 20% weight

Education&Experience: Court

- Court decision
  "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."

Education&Experience: Court

- Court decision
- "limited E&E component"
- "essentially the same today as it was 50 years ago"
- "HRD did not in fact capture these skills in the E&E component."

Education&Experience: Court

- Court decision
  "the effective weight of E&E component is substantially lower than 20% because of the way HRD scores E&E."
- "the final scores on HRD's exams correlated in a perfect linear relationship with the score on the multiple choice tests"

Education&Experience: Court

- Court decision
  "No empirical support or credible professional study justified the 20% weighting."

Validity: Defense

- Test outline based on job analysis
- MC questions covered essential areas including law
- M/C questions measured situational judgment, interpersonal skills, ability to plan, reach logical conclusions based on information at hand.

Validity: Plaintiffs

- Major basis for exam was 1991 study: old
- Newer, Boston job analysis was flawed
  - **Manipulated**

Validity: Plaintiffs

- Major flaws in job analyses
- Illogical ratings of tasks
  - Qualify/practice with weapons daily
  - Talks with leaders of demonstrations daily
- Impossible level of agreement in ratings
  - 26,000 ratings by 11 SMEs with not a single disagreement as to tasks or KSAPs for their own, diverse job assignments
  - Despite different job assignments

Validity: Plaintiffs

- Items tested **rote memory**, not application
- Important KSAPs not tested
- "Easy" items chosen over other items
- SME item review was flawed
- Weights arbitrary
- Passing points arbitrary
- Validity evidence does not support ranking

Validity: Court

- Court decision
  "the SMEs ranked the importance of various KSAs as part of that job analysis. Their rankings are implausible. The 11 SMEs gave identical rankings to all of the approximately 1,100 ratings"

Validity: Court

- Court decision
- "job analysis also claimed that police sergeants perform certain tasks every day, but that could not possibly be true"
  - Practice in operation of firearms/weapons
  - Set up command post
  - Directs major incidents
  - Inspect licensed premises

Validity: Court

- Court decision
  "**With so few [2] SMEs** [reviewing items] and given the deficiencies identified … the **court gives only modest weight to the SME** process in assessing the validity of the exams for statewide application or use in Boston."

Validity: Court

- Court
  "many questions are **definitional** in that the answers turn upon the meaning of a particular word. Those questions have **low fidelity**, because a sergeant's job does not generally involve using academic jargon or other definitions of concepts in the assigned reading.

Validity: Court

- Court decision
  "inability of a written multiple-choice exam to predict good job performance as a sergeant."
- "questions on the exam **largely test for rote memorization** of facts and passages"

Validity: Court

- Court decision
  "serious flaws in identifying (1) which KSAs are testable on a multiple-choice exam and (2) which KSAs are measured in HRD's education and experience component."

Validity: Court

- Court decision
  "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**"

Validity: Court

- Court decision
  "**Chief among the essential skills** … are:
  Leadership skills,
  Supervision skills,
  Decision-making and problem-solving,
  Interpersonal skills,
  Communication skills, and
  Integrity."

Validity: Court

- Court decision
  "According to CP, the most critical determinant of future success as a community policing Officer is:
  A. superior communication skills.
  B. empathy. [key-does not measure empathy]
  C. autonomy.
  D. analytical ability."

Validity: Court

- Court decision
  "make more sense to ask the sergeant candidate what's good practice … as opposed to what's legal"
  "information ... unrelated to the sergeant's job ... the maximum length of prison sentences allowed by law for certain offenses"

Validity: Court

- Court decision
  "[HRD] failed to test meaningfully the KSAs required for good performance as a police sergeant."
- "There is no credible evidence that [the exams] evaluated which information police sergeants must memorize in order to perform their job."

Ranking: Defense

- Ranking required by state law
- Exams measure **some essential** KSAPs
- Exams measure much of job
  - 40% of the KSAs could be tested by written test

Ranking: Plaintiffs

- Ranks imprecise predictors of job perf.
  - Not measure enough of job to rely on ranks
  - Other KSAPs drive job performance also
- Only 22% of KSAPs measured by exams

Ranking: Court

- Court decision
  "No credible study showed that single-point differences in scores reflected any significant difference in job qualifications."
  "HRD itself ... proposed 'banding' ... has already conceded that its multiple-choice examinations were not sufficiently valid as rank order devices, even though they now claim just the opposite."

Pass Point: Defense

- Passing points not important w low selection ratio (i.e., in large PD)
  - Never reach people near the passing point

Pass Point: Plaintiffs

- No attempt to link passing point to job
  - Many easy items
- Passing does not indicate competence
  - No satisfy state law: identify qualified people
- Incumbent sergeants fail the test for sgt.
- Passing point not job related

Pass Point: Court

- Court decision
  "did not rely on any accepted scientific criteria for establishing the passing score for its exams."
- "incumbent sergeants who take HRD promotional exams for lieutenant often do not perform well on the sergeants' portion."

Alternatives: Defense

- Alternatives can increase AI
- Alternatives do not guarantee less AI

Alternatives: Plaintiffs

- Expert's firm pushes alternatives
- In general, alternatives improve validity and reduce AI

Alternatives: Court

- Court decision
  "HRD knew of clearly superior assessment methods, but continued to use the same, unnecessarily discriminatory format anyway."

Alternatives: Court

- Court
  "As [Defense's expert] acknowledged at trial, performance review systems "can be useful and they do tend to reduce adverse impact." His own company … has recommended use of such systems."

Three Related Cases

- Tatum – Police Sergeant
- Lopez – Police Sergeant
- Smith – Police Lieutenant
- Involved similar exams in Massachusetts
- 80% MC, 20% E&E

Dates and Outcomes

- Tatum decision, 2022 – for Plaintiffs
- Smith decision, 2015 – for Plaintiffs
- Lopez decision, 2014 – for Defense
- Similar facts and legal arguments

Lopez 2014

- Court ruled exam was valid despite AI
  - Why?
- Expert for Defense:
  - MC alone not valid
  - With E&E the exam was "minimally valid"
  - No guarantee alternatives would have less AI

Smith 2015

- Exam measured too little of job

Learning Objective 1

- List and describe the major findings of the court decision in Tatum (2022)
- Intentionally used exam type known to have severe AI
- Absurd claims of KSAPs tested by E&E
- One job analysis was clearly manipulated
- Test invalid, not adequate for ranking

Learning Objective 2

- List and describe the positions of the Plaintiffs and the Defense concerning test validity in Tatum (2022)
- D: JK and E&E sufficient for ranking
- P: JK not measure K as used on job
- P: E&E very narrow, missed KSAPs
- P: Highest ranked candidates may be low in essential, unmeasured KSAPs

Learning Objective 3

- List and describe the positions of the Plaintiffs and the Defense concerning adverse impact in Tatum (2022)
- D: Use shift of 1 repeatedly = No AI
- D: Must not aggregate over PDs or years
- D: Require ANOVA before M-W test
- D: Omit PDs with no promotions
- D: Simpson's paradox may be operating

Learning Objective 3

- List and describe the positions of the Plaintiffs and the Defense concerning adverse impact in Tatum (2022)
- P: Aggregation useful to see big picture
- P: Mean differences help interpret AI ratios
- P: Include all PDs in tests of means
- P: No data showing Simpson's Paradox

Takeaways

- Hard to write items on problem solving, especially without good source material
- Study material for law often written by lawyers and includes much extraneous info such as legislative history and penalties (possible length of sentence)
- Must push back at illogical SME ratings

Takeaways

- Be honest in reports and testimony
- Strong arguments better than just plausible
- Judge wanted more than MC/rote memory
- SMEs can be manipulated
- SMEs can make silly ratings
- SMEs like status quo (components/weights)

Related Presentations

- Ways to reduce AI and improve job perf.
- **http://jpwphd.com/ipac2023**

References

- Lopez v. City of Lawrence, U.S. Dist. Ct. No. 07-11693, 2014

Questions from Attendees

- Call me anytime to discuss any of this.
- (617) 244-8859
- jpwiesen@gmail.com