Deadly Statistics:
Quantifying an "Unacceptable Risk" in Capital Punishment

David H. Kaye*

Law, Probability & Risk
Vol. 15, No. 4, Dec. 2016 (in press)

Abstract: In *Atkins v. Virginia*, the U.S. Supreme Court held that the Eighth Amendment ban on cruel and unusual punishment precludes capital punishment for intellectually disabled offenders. Death-penalty states responded with laws defining intellectual disability in various ways. In *Hall v. Florida*, the Court narrowly struck down the use of a measured IQ of 70 to mark the upper limit of intellectual disability because it created "an unacceptable risk that persons with intellectual disability will be executed." But the Court was unclear if not inconsistent in its description of an upper limit that would be acceptable. Four dissenting Justices accused the majority not only of misconstruing the Eighth Amendment, but also of misunderstanding elementary statistics and psychometrics. This article uses more complete statistical reasoning to explicate the Court's concept of unacceptable risk. It describes better ways to control the risk of error than the Court's confidence intervals, and it argues that, to the extent that the Eighth Amendment allows any quantitative cut score in determining an offender's intellectual disability, these more technically appropriate methods are constitutionally permissible.

Keywords: *Hall v. Florida*, cruel and unusual, Eighth Amendment, capital punishment, intellectual disability, IQ, psychometrics, cut-score, measurement error, standard error, confidence interval, shrinkage estimator, Bayesian inference, credible region, burden of persuasion

* Associate Dean for Research, Distinguished Professor, and Weiss Family Scholar, The Pennsylvania State University, Penn State Law. Preliminary versions of this article were presented at the Seventh International Conference on Inference and Forensic Statistics, Leiden, August 2014, a Penn State Law Faculty Colloquium, January 2015, and the Joint Statistical Meetings, Seattle, August 2015. I am grateful to Johannes Fredderke, Jay Kadane, Mae Quinn, and an anonymous referee for their comments on a draft and to Jim Greiner, Jay Kadane, Jay Koehler, and Hal Stern for comments on a related paper.

Deadly Statistics:
Quantifying an "Unacceptable Risk" in Capital Punishment

Contents

Introduction

> "The record of how human beings have performed on IQ tests does more than measure us against one another for entry into Harvard law school." — JAMES R. FLYNN, INTELLIGENCE AND HUMAN PROGRESS 3 (2013)

For some of us, a single IQ point can mark the difference between life and death. Such is the teaching of *Hall v. Florida*.[1] *Hall* is the Supreme Court's third — and most detailed — case to address just when a criminal who might be intellectually disabled can nevertheless be sentenced to death. In *Hall*, the Court held unconstitutional Florida's use of a measured IQ score of 70 to mark the upper limit of intellectual disability. Four dissenting Justices were incensed. In an opinion written by Justice Alito,[2] they interpreted Justice Kennedy's majority opinion[3] to mean that only IQ scores of 75 or more could serve as a cutoff.[4] Using their own computations of "confidence intervals,"[5] they concluded that this rule "totally transforms the allocation and nature of the burden of proof."[6] The Court reached a rule "unhinged from legal logic,"[7] Justice Alito lambasted, because it "misunderstands how the SEM [standard error of measurement] works"[8] and made "factual mistakes that will surely confuse States attempting to comply with its opinion."[9]

The opinions of this divided Court have escaped probing analysis of their treatment of the statistical issues in ascertaining intellectual disability.[10] Impressed with potential doctrinal implications of the case for criminal jurisprudence more generally, commentators have seen it as

---

[1] 134 S.Ct. 1986 (2014).

[2] Chief Justice Roberts and Justices Scalia and Thomas joined Justice Alito's opinion. *Id*. at 2001.

[3] Justices Breyer, Ginsburg, Sotomayor, and Kagan joined Justice Kennedy's opinion.

[4] *Id*. at 2003, 2010 (Alito, J., dissenting). *But see id*. at 2011 ("it is unclear to me" which rule the Court adopts).

[5] *Id*. at 2010–11.

[6] *Id*. at 2011.

[7] *Id*.

[8] *Id*. at 2009.

[9] *Id*. at 2010.

[10] Most commentary stops with the recognition that "standardized measures will always be imprecise," Susan Unok Marks, *Courts' Elusive Search for the Meaning of Intellectual Disability for Evaluating* Atkins *Claims*, 26 U. FLA. J.L. & PUB. POL'Y 347 (2015), or demands that "any IQ score be reported with an associated confidence interval." Robert M. Sanger, *IQ, Intelligence Tests, "Ethnic Adjustments" and* Atkins, 65 AM. U. L. REV. 87 (2015). *But see* David H. Kaye, *"Unhinged from Legal Logic":* Hall v. Florida*'s Confidence Intervals and the Burden of Persuasion* (Aug. 15, 2015) (unpublished manuscript).

supporting challenges to other aspects of capital sentencing procedures.[11] Even more broadly, one astute scholar of mental health law sensed in *Hall* the seeds of a general "scientization of the criminal law."[12] This commentary on the implications of *Hall* is intriguing and important, but it leaves undisturbed and unexamined the statistical science in *Hall* and its relationship to constitutional constraints and legal policies.

Therefore this article focuses on the statistical issues in the case. It examines the more important statements in the Justices' opinions on measurement error in IQ testing. These statements are of more than didactic and technical interest. The statistical properties of IQ scores are critical to understanding the capital sentencing options retained by states after *Hall*. The *Hall* majority condemned Florida for ignoring "a statistical fact"[13] that constituted "one of the most important concepts in measurement theory."[14] Without careful analysis of the relevant statistical principles, it is not possible to ascertain how much *Hall* should constrain states that wish to carve out a range of IQ scores that would preclude further proof of intellectual disability. Furthermore, because IQ scores are of major importance in the broader clinical assessments that courts must review, understanding their statistical properties and limitations is vital for all adjudication of intellectual-disability claims.

Part I of this article sets the stage for the examination of the exchange among the Justices on the technical concepts such as SEM and confidence intervals. It describes the issue in *Hall* as it emerged in the Court's "zig-zagging death penalty jurisprudence"[15] and the general divide between the majority and the dissent. It reads the majority opinion as accepting the premise that a state may, without further inquiry, deem all individuals with true IQs above a fixed number as not intellectually disabled but that it must adopt some margin of error in applying this categorical rule to measured scores.

Part II explains why the state's choice of an IQ cut-score, whether conceived of as a true score or framed as a measured one, is inherently arbitrary. It shows how the measured score of 70

---

[11] *See* Timothy R. Saviello, *The Appropriate Standard of Proof for Determining Intellectual Disability in Capital Cases: How High Is Too High?*, 20 BERKELEY J. CRIM. L. 163 (2015) (arguing that *Hall* indicates that the state cannot impose a greater burden on the defendant than proving intellectual disability by a preponderance of the evidence); *Eighth Amendment—Cruel and Unusual Punishments—Defendants with Intellectual Disability*—Hall v. Florida, 128 HARV. L. REV. 271, 280 (2014) (suggesting that *Hall* invites attacks on all death penalty procedures that are legislative outliers; on the significance of various types of legislative outliers in constitution al litigation more generally, see Justin Driver, *Constitutional Outliers*, 81 U. CHI. L. REV. 929 (2014)); Lise E. Rahdert, Hall v. Florida *and Ending the Death Penalty for Severely Mentally Ill Defendants*, 124 YALE L.J. FORUM 34 (2014) (proposing that the case can readily be extended to prohibit executions of every criminal suffering from a major mental illness).

[12] Christopher Slobogin, *Scientizing Culpability: the Implications of* Hall v. Florida *and the Possibility of a "Scientific Stare Decisis,"* 23 WM. & MARY BILL RTS. J. 415, 425 (2014).

[13] *Hall*, 134 S.Ct. at 1995.

[14] *Id*.

[15] Scott E. Sundby, *The True Legacy of* Atkins *and* Roper*: The Unreliability Principle, Mentally Ill Defendants, and the Death Penalty's Unraveling*, 23 WM. & MARY BILL RTS. J. 487, 493 (2014).

selected by Florida as well as the not-fully-specified but higher numbers demanded by the majority affect the proportion of the population who could be found to be intellectually disabled.

Part III analyzes the standard error of measurement and resulting confidence intervals that are central to the Court's disposition of the case. It clarifies ambiguities in the Court's opinion and criticizes the dissent's effort to link a confidence interval with the burden of persuasion. It notes that the statistical apparatus of confidence intervals is not even necessary to establish a cutoff for observed IQ scores that accounts for the SEM. Finally, this Part identifies a different statistic — the standard error of estimate (SEE) — that is more appropriate than the Court's SEM in adjusting for measurement error.

Part IV moves beyond the situation of a single IQ score and the confines of classical test theory. First, it considers the Justices' discussion of combining IQ scores to achieve greater precision. It shows that the majority's comments on the difficulty of aggregation should not bar the use of established methods. Similarly, it suggests that the exclusive reliance on classical test theory and confidence intervals in the opinions should not preclude the use of Bayesian statistical procedures that would permit a more perspicacious treatment of the problem of measurement error in IQ scores than the frequentist one that the Court endorsed.

I. The Intellectual Disability Trilogy

In the space of 25 years, the Supreme Court has moved from permitting the state to execute a man with an IQ in the 50–63 range[16] to barring the state from executing a man with an IQ in the 71–80 range.[17] This transformation began in 1989, in *Penry v. Lynaugh*.[18] A welter of conflicting opinions in *Penry* established that before imposing a capital sentence, the sentencing judge or jury must have the opportunity to consider and give appropriate weight[19] to what was then called mental retardation.[20] At the same time, the Court declined to hold that the Eighth Amendment's ban on cruel

---

[16] Penry v. Lynaugh, 492 U.S. 302, 307 (1989).

[17] Hall v. Florida, 134 S.Ct. 1986, 1992 (2014). Earlier IQ scores as low as 60 were ruled inadmissible. *Id*.

[18] 492 U.S. 302 (1989), *overruled in part*, Atkins v. Virginia, 536 U.S. 304 (2002).

[19] A majority of the Court held that the issues as framed in a special verdict for the jury did not allow proper consideration of Penry's intellectual disability. Justice Scalia's opinion, joined by Chief Justice Rehnquist and Justices White and Kennedy, dissented from this conclusion. *Penry*, 492 U.S. at 353–60 (dissenting and concurring opinion).

[20] For many years, "mental retardation" was the phrase that the legal, medical, and psychological communities used to denote certain deficits in cognitive functioning. *See* AM. PSYCHIATRIC ASS'N, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS 33 (5th ed. 2013) [hereinafter DSM-5]; Robert L. Schalock et al., *The Renaming of Mental Retardation: Understanding the Change to the Term Intellectual Disability*, 45 INTELL. & DEVELOPMENTAL DISABILITIES 116, 116 (2007). This essay generally uses what one court has called "the more politically correct phrase 'intellectual disability.'" Commonwealth v. Hackett, 99 A.3d 11, 14 n.5 (Pa. 2014). In some instances where it is historically appropriate, however, I use the earlier terminology. *Cf*. PATRICK MCDONAGH, IDIOCY: A CULTURAL HISTORY 5 (2008)

and unusual punishment insulated *all* mentally retarded individuals from death at the hands of the state. Writing for the majority, Justice O'Connor concluded that "there is insufficient evidence of a national consensus against executing mentally retarded people convicted of capital offenses for us to conclude that it is categorically prohibited by the Eighth Amendment."[21] Four dissenting Justices did not dispute this assessment of national opinion, but they insisted that capital punishment is unacceptable for "the mentally retarded as a class"[22] because their limitations make the death sentence invariably disproportionate.[23]

A scant thirteen years later, in *Atkins v. Virginia*,[24] the Court found a "dramatic shift in the state legislative landscape."[25] The requisite "national consensus"[26] that it is categorically wrong to execute "the mentally retarded" had coalesced.[27] Moreover, Justice Stevens's majority opinion buttressed this conclusion with the arguments of the dissenters in *Penry*. The execution of the mentally retarded, he maintained, does not measurably advance the goals of retribution and

---

("[a]nyone wanting to understand the history of the idea of intellectual disability and its various genealogical precursors, such as idiocy, must contend with the slipperiness of the key terms").

[21] *Penry*, 492 U.S. at 335.

[22] *Id*. at 343 (Brennan & Marshall, JJ., concurring and dissenting).

[23] Justices Brennan and Marshall contended that the mentally retarded "inevitably lack the cognitive, volitional, and moral capacity to act with the degree of culpability associated with the death penalty," *id*. at 343–44, and that "killing mentally retarded offenders does not measurably further the penal goals of either retribution or deterrence." *Id*. at 348. Justices Stevens and Blackmun apparently agreed. *See id*. at 350. Justice O'Connor did not. *Id*. at 338. The remaining four Justices dismissed the entire inquiry as misguided. In their view, "if an objective examination of laws and jury determinations fails to demonstrate society's disapproval of it, the punishment is not unconstitutional even if out of accord with the theories of penology favored by the Justices of this Court." *Id*. at 351.

A different criticism of *Penry* is that the Court failed to appreciate that the English and colonial common law protected not "only those who were 'profoundly or severely retarded,' [but also many of] those who were moderately or mildly mentally retarded." Michael Clemente, Note, *A Reassessment of Common Law Protections for "Idiots,"* 124 YALE L.J. 2746, 2751 (2015).

[24] 536 U.S. 304 (2002).

[25] *Id*. at 310.

[26] *Id*. at 316–17.

[27] *Id*. at 316. A dissenting opinion of Justice Scalia, Chief Justice Rehnquist, and Justice Thomas found this conclusion "miraculously extract[ed]" from "embarrassingly feeble evidence." *Id*. at 342, 344.

deterrence.[28] In addition, he linked the substantive conclusion that it is wrong to execute anyone who is intellectually disabled to the accuracy of the procedures for identifying death-worthy defendants:

> The reduced capacity of mentally retarded offenders provides a second justification for a categorical rule making such offenders ineligible for the death penalty. The risk that the death penalty will be imposed in spite of factors which may call for a less severe penalty . . . is enhanced, not only by the possibility of false confessions, but also by the lesser ability of mentally retarded defendants to make a persuasive showing of mitigation in the face of prosecutorial evidence of one or more aggravating factors. Mentally retarded defendants may be less able to give meaningful assistance to their counsel and are typically poor witnesses, and their demeanor may create an unwarranted impression of lack of remorse for their crimes. As *Penry* demonstrated, moreover, reliance on mental retardation as a mitigating factor can be a two-edged sword that may enhance the likelihood that the aggravating factor of future dangerousness will be found by the jury. Mentally retarded defendants in the aggregate face a special risk of wrongful execution.[29]

But *Atkins* left unclear and unresolved the question of precisely who is "mentally retarded" for Eighth Amendment purposes. Because *Penry* permitted capital punishment of unequivocally intellectually disabled offenders, it created little pressure to define intellectual disability carefully. By fashioning a blanket exemption from capital punishment for everyone "within the range of mentally retarded offenders about whom there is a national consensus,"[30] however, *Atkins* brought the problem of line-drawing to the foreground. Yet, the Court made no effort to define "the range of mentally retarded offenders." Instead, it grandly announced that "in determining which offenders are in fact retarded . . . 'we leave to the State(s) the task of developing appropriate ways to enforce the constitutional restriction . . . .'"[31]

Still, *Atkins* dropped some breadcrumbs. A footnote implied that the "clinical definitions" of the American Association on Mental Retardation (AAMR) and the American Psychiatric

---

[28] *Id*. at 319–20. However, whereas the *Penry* dissenters offered this as a sufficient basis for a determination of cruel and unusual punishment, the *Atkins* majority proffered it only insofar as "independent evaluation of the issue reveals no reason to disagree with the judgment of the legislatures that have recently addressed the matter and concluded that death is not a suitable punishment for a mentally retarded criminal." *Id*. at 320 (internal quotation marks omitted). Justice Scalia responded that the "discussion . . . does not bear analysis." *Id*. at 350.

[29] *Id*. at 320–21 (internal quotation marks and citation omitted). Justice Scalia scorned "this unsupported claim" and its "pretty flabby language." *Id*. at 352. For a sympathetic exploration of the radical implications of this "unreliability principle" as a basis for holding capital sentences to be cruel and unusual, see Sundby, *supra* note 15.

[30] *Id*. at 317.

[31] *Id*. (quoting Ford v. Wainwright, 477 U.S. 399, 405, 416–17 (1986)).

Association (APA) would do the trick.[32] The AAMR's definition referred to "significantly subaverage intellectual functioning, existing concurrently with related limitations in two or more . . . adaptive skill areas . . . and manifest[ing] before age 18."[33] The fourth edition of the APA's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-4) was "similar" and added that "'[m]ild' mental retardation is typically used to describe people with an IQ level of 50-55 to approximately 70."[34] In addition, in concluding that execution of the mentally retarded had become "truly usual,"[35] the Court observed that "only five [states] have executed offenders possessing a known IQ less than 70 since we decided *Penry*."[36]

States responded to *Atkins*' Olympian command or to earlier public discomfort with executions of the intellectually disabled[37] in various ways. Several states enacted laws using a measured IQ score of 70 (or a corresponding number of standard deviations below the population mean) as the upper limit for intellectual disability.[38] Florida had enacted such a law even before *Atkins*. Its statute, as interpreted by the state's supreme court, effectively "defined intellectual disability to require an IQ test score of 70 or less. If, from test scores, a prisoner is deemed to have an IQ above 70, all further exploration of intellectual disability is foreclosed."[39]

Florida applied its statute to Robert Lee Hall. Although Hall's family and educational history provided distressing evidence of intellectual disability,[40] the IQ scores deemed admissible at a post-*Atkins* hearing were too high to permit this compelling evidence to be considered. They ranged from 71 to 80.[41] And so, twelve years after *Atkins*, the Court squarely confronted the issue of defining

---

[32] *Id*. at 317 n.22 (The statutory definitions of mental retardation are not identical, but generally conform to the clinical definitions . . . .").

[33] *Id*. at 308 n. 3 (quoting AMERICAN ASSOCIATION ON MENTAL RETARDATION, MENTAL RETARDATION: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS 5 (9th ed. 1992).

[34] *Id*. (quoting DSM-4, at 42–43).

[35] *Id*. at 316.

[36] *Id*.

[37] *See id*. at 314.

[38] *Hall*, 134 S.Ct. at 1996. Kentucky's statute defined "significantly subaverage general intellectual functioning . . . as an intelligence quotient (I.Q.) of seventy (70) or below." Ky. Rev. Stat. § 532.130(2) (upheld in Bowling v. Commonwealth, 163 S.W. 3d 361 (Ky. 2005)). Virginia required "performance on a standardized measure of intellectual functioning administered in conformity with accepted professional practice, that is at least two standard deviations below the mean." Va. Code Ann. § 19.2–264.3:1.1.

[39] *Id*. at 1990. Like Virginia, Florida defined "significantly subaverage general intellectual functioning" as "performance that is two or more standard deviations from the mean score on a standardized intelligence test." Fla. Stat. § 921.137(1) (2002) (interpreted in *Cherry v. State*, 959 So.2d 702, 712–713 (Fla. 2007) to denote a score of 70 on an IQ test with a mean of 100 and a standard deviation of 15).

[40] *Id*. at 1990–91; Hall. v. Florida, 614 So.2d 473, 479, 479–80 (1993) (dissenting opinion).

[41] *Hall*, 134 S.Ct. at 1992.

intellectual disability in *Hall v. Florida*. Descending from Olympus to Delphi, Justice Kennedy invoked the APA's view as expressed in its *Diagnostic and Statistical Manual* (DSM) that because IQ scores are imprecise, a somewhat higher measured score than 70 could justify clinical inquiry into the individual's adaptive functioning in school, in the family, and elsewhere.[42] Thus, professional criteria that had appeared in *Atkins* to be merely sufficient for a finding of intellectual disability — because they were roughly congruent with legislative and popular views of what it means to be so disabled — turned out to be "a fundamental premise of *Atkins*."[43]

It is easy to read this statement in *Hall* about the DSM as declaring that this branch of death penalty jurisprudence must track the diagnostic criteria of psychiatrists and psychologists.[44] Adopting this reading of *Hall*, the dissent scoffed that the majority had equated "evolving standards of decency" — which is the traditional test for determining the scope of the Eighth Amendment[45] — with "the evolving standards of professional societies."[46] But the majority's language is more restrained. Justice Kennedy cautioned that the professional standards informed but did not dictate the outcome.[47] Even so, the Court gave great weight to the views of the psychiatrists and psychologists. In this case at least,[48] *their* standards informed the Court not only that there was a unanimous expert consensus,[49] but that the consensus rested on a "statistical fact" — "one of the most important concepts in measurement theory"[50] — that "[e]ach IQ test has a 'standard error of measurement,' . . . often referred to by the abbreviation 'SEM.'"[51] And, most other capital punishment states accepted the professionally prevalent practice of considering adaptive functioning for individuals with IQ scores as high as two SEMs above 70.[52] From these facts alone, the Court

---

[42] *Id*. at 1998–99.

[43] *Id*. at 1999.

[44] *See id*. at 2001 (dissenting opinion); *see also* Slobogin, *supra* note 12, at 416.

[45] Trop v. Dulles, 356 U.S. 86, 101 (1958).

[46] *Hall*, 134 S.Ct. at 2002 (Alito, J., dissenting).

[47] Justice Kennedy explained that "[i]n addition to the views of the States and the Court's precedent, [our] determination is informed by the views of medical experts. These views do not dictate the Court's decision, yet the Court does not disregard these informed assessments." *Id*. at 2000.

[48] *Id*. ("the professional community's teachings are *of particular help in this case*, where no alternative definition of intellectual disability is presented and where this Court and the States have placed substantial reliance on the expertise of the medical profession.") (emphasis added).

[49] *See id*. at 2000 (relying on "the unanimous professional consensus").

[50] *Id*. at 1995.

[51] *Id*.

[52] *Id*. at 1996–97.

inferred that Florida's "rigid rule . . . creates an unacceptable risk that persons with intellectual disability will be executed and thus is unconstitutional."[53]

Strictly speaking, this conclusion — that a state may not use a measured-score cutoff as low as 70 — is as far as the Court's holding goes. This holding is grounded upon the concept of a "true score." Intuitively, the true score is the defendant's actual IQ, which might be somewhat different from each and every measurement of it.[54] The case establishes that (when it is consistent with the unanimous position of the mental health community)[55] a state may conclusively presume that all defendants whose IQs *truly* are above a specific number (such as 70) are not intellectually disabled — but it may not apply this presumption to defendants with *measured* scores that are only somewhat higher than this targeted number. This holding invites a series of questions: Why can *any* true score be the basis for a *conclusive* presumption of nondisability? How low can this score be? How much higher than the lowest true-score cutoff must the lowest *measured* cut-score be? Part II responds to the first two questions. Parts III and IV address the last question.

II. The Need to Allow States to Use Cut Scores and the Meaning of "Significantly Subaverage"

The *Hall* Court did *not* reject all IQ cutoffs as unconstitutionally rigid.[56] One could imagine an opinion striking down the Florida law by asserting that intellectual disability is a single, overarching category to be managed by psychologists or psychiatrists informed by multiple streams of information about the constructs of both intellectual and adaptive functioning, and holding that

---

[53] *Id*. at 1999.

[54] Part III provides the technical definition.

[55] *See Hall*, 134 S.Ct. at 2000 ("By failing to take into account the SEM and setting a strict cutoff at 70, Florida goes against the unanimous professional consensus.") (internal quotation marks omitted). Whether all mental health professionals agree that the cutoff should be higher than 70 is doubtful. *See* Bryan H. King et al., *Mental Retardation*, *in* 2 KAPLAN AND SADOCK'S COMPREHENSIVE TEXTBOOK OF PSYCHIATRY 2587, 2591 (Benjamin J. Sadock & Virginia A. Sadock eds., 7th ed. 2000) (referring to extending "the I.Q. criterion from '70 and below' to '70 or 75 and below'" as having been "hotly debated").

[56] Some commentary conveys a different impression. *E.g.*, Sanger, *supra* note 10, at 93 (attributing the Court the "conclusion that rigid reliance on IQ scores should not deprive people facing the death penalty of a chance to illustrate that their execution is unconstitutional"); Timothy R. Saviello, *The Appropriate Standard of Proof for Determining Intellectual Disability in Capital Cases: How High Is Too High?*, 20 BERKELEY J. CRIM. L. 163, 222 (2015) (reading *Hall* as resting on the premise that "having a fixed IQ cutoff makes the IQ score the single criteria for determining intellectual disability, and thus prevents consideration of other evidence that mental health professionals require prior to reaching a decision on intellectual disability"); Bryant Buechele, Note, *Psychology's Role in Law: A Discussion of How the Supreme Court Views the Role of the DSM-V in* Hall v. Florida, 68 SMU L. REV. 275, 277 (2015) (because "IQ score and adaptive functioning, by themselves are not capable of determining whether an individual is intellectually disabled, . . . *Hall* found that a Florida statute that effectively did not account for factors beyond the somewhat faulty IQ test was unconstitutionally limited in scope.").

the Florida law was unconstitutional because it did not allow these experts to present all that information. This outcome would have been even more deferential to the expert community than was Justice Kennedy's opinion in *Hall*. It would have resolved the matter by decisively committing every *Atkins* claim to warmer and fuzzier clinical judgments as opposed to erecting a cold and unyielding statistical barrier.[57]

But could the Eighth Amendment (or any other part of the Constitution) be held to prevent a state from using *some* cut-score for administrative convenience? Surely, at some point a high IQ score means, *ipso facto*, there is no significant risk of executing someone who cannot be said to deserve this punishment because of an intellectual incapacity. The "unacceptable risk" approach — which implies that some level of risk is acceptable — was a less radical response to implementing *Atkins*. The *Hall* majority did not dispute Justice Alito's observation that "[a] defendant who does not display significantly subaverage intellectual functioning is [simply] not among the class of defendants we identified in *Atkins*."[58] "Significantly subaverage intellectual functioning" as introduced in *Atkins* is integral to the conceptualization of intellectual disability in psychology, and a standardized test is a more objective and reliable instrument than is a clinician's impressions of how far a defendant's intellect is from the population norm. Medically, a substantial impairment in

[57] On the relative performance of statistical methods and clinical judgment more generally, see R.M. Dawes et al., *Clinical Versus Actuarial Judgment*, 243 SCIENCE 1668 (1989) ("Research comparing these two approaches shows the actuarial method to be superior."); William M. Grove, & Paul E. Meehl, *Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–statistical Controversy*, 2 PSYCH., PUB. POL'Y & L. 293 (1996) ("Empirical comparisons of the accuracy of the two methods (136 studies over a wide range of predictands) show that the mechanical method is almost invariably equal to or superior to the clinical method); Konstantinos V. Katsikopoulos et al., *From Meehl to Fast and Frugal Heuristics (and Back): New Insights into How to Bridge the Clinical—Actuarial Divide*, 18 THEORY & PSYCH. 443 (2008); STEVEN SCHWARTZ & TIMOTHY GRIFFIN, MEDICAL THINKING: THE PSYCHOLOGY OF MEDICAL JUDGMENT AND DECISION MAKING (2012).

[58] *Id*. at 2009 (Alito, J., dissenting). The phrase appears in *Atkins*, 536 U.S. at 308 n. 3. The *Atkins* Court took it from APA and AAMR publications. *Id*.

intellect is necessary to a diagnosis of intellectual disability,[59] and low IQ scores currently are essential to an expert determination of a substantial impairment.[60]

IQ scores are so important because intellectual disability is not a condition that can be attributed to a single mechanism such as the loss of dopamine-producing brain cells that causes Parkinson's.[61] It is no less real,[62] but its very definition is based in large part on its rarity in the

---

[59] Justice Alito asserted that in the latest edition of its *Diagnostic and Statistical Manual* (the DSM-5), "the APA discards 'significantly subaverage intellectual functioning' as an element of the intellectual-disability test. Elevating the APA's current views to constitutional significance therefore throws into question the basic approach that *Atkins* approved and that most of the States have followed."). Examination of the references in an accompanying footnote does not support this assertion. Subaverage intellectual functioning, as measured by low IQ scores, remains the first diagnostic criterion. *See* David H. Kaye, *Quarreling and Quibbling over Psychometrics in* Hall v. Florida *(part 3)* (June 4, 2014), FORENSIC SCI., STAT. & L. http://for-sci-law-now.blogspot.com/2014/06/quarreling-and-quibbling-over_4.html (concluding that Justice Alito's "claim that the [DSM-V] 'dramatically illustrate[s a] fundamental[] alter[ation in] ... the longstanding ... definition of intellectual disability' seems, well, melodramatic"); APA, DSM-5 Intellectual Disability Fact Sheet, at 2 (2013), available at http://www.dsm5.org/Documents/Intellectual%20Disability%20Fact%20Sheet.pdf ("In DSM-5, intellectual disability is considered to be approximately two standard deviations or more below the population, which equals an IQ score of about 70 or below.").

[60] The APA's DSM-5 demands "[d]eficits in intellectual functions . . . confirmed by both clinical assessment *and individualized, standardized intelligence testing*." DSM-5, *supra* note 20 (Diagnostic Criterion A) (emphasis added). Likewise, the American Association on Intellectual and Developmental Disabilities (AAIDD) defines intellectual disability in terms of "significant limitations *both* in intellectual functioning and adaptive behaviour." AAIDD, INTELLECTUAL DISABILITY: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS (11th ed. 2010) (as quoted in CHRIS HATTON, INTELLECTUAL DISABILITIES—CLASSIFICATION, EPIDEMIOLOGY AND CAUSES IN CLINICAL PSYCHOLOGY AND PEOPLE WITH INTELLECTUAL DISABILITIES 1, 4 (Eric Emerson et al. eds., 2d ed. 2012) (emphasis added). The organization then defines "intellectual functioning" as "[a]n IQ score that is approximately two standard deviations below the mean, considering the standard error of measurement for the specific assessment instruments' used and the instruments' strengths and limitations." *Id*.

[61] The more extreme and debilitating levels of disability are the product of known organic causes and occur more frequently than would be predicted from the normal curve. The severe cases elevate and fatten the tails of the distribution of scores. The much larger fraction of cases (perhaps 90%) probably result from interactions between quantitative trait loci that influence intellectual development in combination with environmental conditions. These cases would be expected to be normally distributed in the population. 2 KAPLAN & SADOCK'S COMPREHENSIVE TEXTBOOK OF PSYCHIATRY 2592 (Benjamin J. Sadock & Virgina A. Sadock eds., 7th ed. 2000).

[62] Many of the diagnostic categories in the latest version of the Diagnostic and Statistical Manual of Mental Disorders have been questioned. The DSM-5 "was published in May 2013 amid a storm of controversy and bitter criticism." News Analysis: Controversial Mental Health Guide DSM-5, NHS Choices (Aug. 15, 2013) http://www.nhs.uk/news/2013/08august/pages/controversy-mental-health-diagnosis-and-treatment-dsm5.aspx. In general, critics maintain that "D.S.M.'s diagnostic categories lacked validity, that they were not 'based on any objective measures,' and that, 'unlike our definitions of ischemic heart disease,

population. The question of how rare is rare enough has plagued efforts to develop satisfactory diagnostic guidelines. Over the last 50 years, proposed IQ cutoff scores have varied from one to two standard deviations (SDs) below the population mean.[63] The popularity of the "more traditional"[64] figure of two SDs seems to reside in a desire to keep the number of diagnoses low, partly to avoid stigmatizing large numbers of minorities in school settings.[65]

As the number of SDs for the cutoff grows, the proportion of people who are subject to the classification shrinks. IQ scores, like height, weight, and many other physical characteristics,[66] tend to have a "normal" or "Gaussian" distribution in the population.[67] Consequently, this relationship is strongly nonlinear in the vicinity of –2 SD. A score change of a few points can dramatically change the proportion of people who are affected. Because this fact has a good deal to do with the professional preference for one cut-score over another, it is worth pausing to draw a picture of where the cut-scores discussed in *Hall* lie.

The Florida legislature defined "significantly subaverage general intellectual functioning" as "performance that is two or more standard deviations from the mean score on a standardized intelligence test."[68] Justice Kennedy translated this into the scale on which IQ scores are reported as

---

lymphoma or AIDS,' which are grounded in biology, they were nothing more than constructs put together by committees of experts." Gary Greenberg, *The Rats of N.I.M.H.*, NEW YORKER, May 16, 2013, available at http://www.newyorker.com/tech/elements/the-rats-of-n-i-m-h (quoting Thomas Insel, the director of the National Institute of Mental Health).

[63] Daniel J. Reschly, *Assessing Mild Intellectual Disability: Issues and Best Practice*, *in* THE OXFORD HANDBOOK OF CHILD PSYCHOLOGICAL ASSESSMENT 683, 687 (Donald H. Saklofske et al. eds., 2013). For a synopsis of the history of definitions, see COMMITTEE ON DISABILITY DETERMINATION FOR MENTAL RETARDATION, NAT'L RESEARCH COUNCIL, MENTAL RETARDATION: DETERMINING ELIGIBILITY FOR SOCIAL SECURITY BENEFITS 22–24 (2002).

[64] Reschly, *supra* note 63, at 687.

[65] The more liberal criterion of one SD "markedly influenced ID criteria in schools and was the subject of much criticism in the courts and by researchers as being too inclusive and stigmatizing excessive numbers of persons." *Id*. (citations omitted); NRC COMMITTEE, *supra* note 63, at 24 (noting that a significant factor in reducing "the upper criterion of scores on intelligence measures from 85 to 70" in 1973 was " concern about the inappropriate overidentification of minority students as mentally retarded"); *cf*. Sadock & Sadock, *supra* note 61, at 2591 (noting the concern that raising the cut score from 70 points (–2 SD) to 75 "will increase the size of the population with mental retardation, including increases in the over-representation of several minority groups").

[66] *See* PETER H. WESTFALL & KEVIN S.S. HENNING, UNDERSTANDING ADVANCED STATISTICAL METHODS 272 (2013) ("Natural processes everywhere are well modeled by the notmal distribution. It's not just a figment of the imagination of a few deranged statisticians.").

[67] On why this might be so, see Aidan Lyon, *Why Are Normal Distributions Normal?*, 65 BRITISH J. PHIL. SCI. 621 (2014).

[68] Cherry v. State, 959 So.2d 702, 712 (Fla. 2007); *cf*. Atkins v. Virginia, 536 U.S. 304, 308 (2002) ("significantly subaverage intellectual functioning").

follows:

> The concept of standard deviation describes how scores are dispersed in a population. . . . The standard deviation on an IQ test is approximately 15 points, and so two standard deviations is approximately 30 points. Thus a test taker who performs "two or more standard deviations from the mean" will score approximately 30 points below the mean on an IQ test, i.e., a score of approximately 70 points.[69]

The "normal" distribution is the bell-shaped one prominent in elementary statistics courses. The exact shape of the family of all normal distributions can be determined from just two numbers — the mean and the standard deviation. The mean states where the bell sits, and the standard deviation determines how steeply its sides flow down from the top.

A few symbols will help immeasurably in keeping track of the different quantities that are important in *Hall*. We can use x to refer to all the IQ scores in a population, μ_x to denote their mean, and σ_x to represent their standard deviation. As is conventional in statistics, z will denote the scores expressed in units of standard deviations.[70] As shown in Figure 1, for an IQ test with a population mean of $\mu_x = 100$ and a population SD of $\sigma_x = 15$, 2.5% of the scores lie at or below 70 ($z = -2$, which is to say $2\sigma_x$ below the mean). More than twice as many, 5.1%, fall at or below 75 ($z = -1.67$), and nearly seven times as many, 16.7%, occur at or below 85 ($z = 1$).

---

[69] *Hall*, 134 S. Ct. at 1994. It would have been more precise to state that the standard deviation indicates not *how*, but rather *how much* scores are dispersed in either a population or a sample. A batch of numbers could be highly concentrated around a single value, with outliers on the flanks. Their distribution could be flat, with an equal fraction of the numbers spread out everywhere on the number line. The distribution might show clustering at several locations.

[70] Z is obtained by subtracting the mean IQ score μ_x from the IQ score x and dividing this departure from the mean by the standard deviation of x. In symbols, $z = (x - \mu_x) / \sigma_x$.

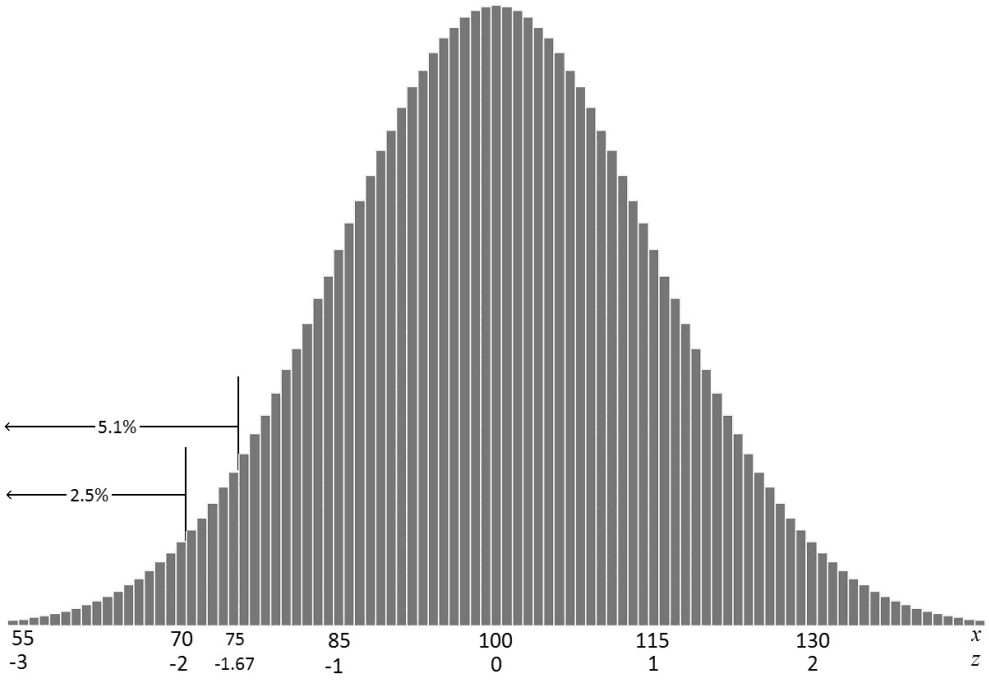

Figure 1. Frequency Distribution of IQ Scores in a Population with a Mean of 100 and a Standard Deviation of 15. Note that 2.5% of the scores lie below 71 (the Florida cutoff), 5.1% are below 76 (the *Hall* Court's suggestion), and 16.7% are below 86 (an older cutoff).[71]

It may be worth summarizing what we have learned about the law, psychology, and statistics of IQ cut scores so far. Both opinions in *Hall* treat a major deficit in intellectual (as opposed to adaptive) functioning as a necessary but not sufficient condition for death-penalty disqualification, and both regard IQ tests as a valid measure of the relevant intellectual functioning.[72] For decades,

[71] IQ scores are integers, and the height of each bar is proportional to the relative number of people at each integer. The normal curve is a continuous line that fits these heights. In this figure and throughout this Essay, I use the area under the curve from $x - 0.5$ to $x + 0.5$ to compute the proportion of people with an integral IQ score x.

[72] Establishing that IQ scores are valid in ascertaining who is eligible for capital punishment requires two steps. To begin with, one needs a list of the factors that make intellectual ability relevant to capital sentencing. These factors flow from the two sets of justifications for prohibiting capital punishment of individuals classified as "intellectually disabled." The first set relate to the judgment that execution of the disabled is morally wrong. *See* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY (1977). In *Hall*, Justice Kennedy reiterated the Brennan-Marshall view that executing a "person with intellectual disability" constitutes cruel and unusual punishment because it serves "[n]o legitmate penological purpose" in that "those with intellectual disability are . . . likely unable to make . . . calculated judgments . . . ." *Hall*, 134 S. Ct. at 1992–93. The second set of justifications pertains to the ways in which intellectual disability degrades the accuracy of sentencers' judgments of which offenders truly deserve to die. The *Hall* majority explained that "intellectually disabled . . . persons face 'a special risk of wrongful execution' because they are more

behavioral scientists and clinicians, seeking to limit the percentage of the population that would be labeled intellectually disabled and appreciating the shape of the normal curve that describes the distribution of IQ in the general population[73] have deemed a departure of at least twice the population standard deviation σ_x as indicative of the "significantly subaverage intellectual functioning" that is essential to a diagnosis of disability.

The choice of any particular cut-score to mark "significantly subaverage" obviously is arbitrary, for there is no precise point at which the quality "significantly subaverage" springs into existence. There is little or no difference in intellectual functioning in people who are within an IQ point or two of one another. Thus, the choice of any particular number of SDs is largely a convention.[74]

The inherent room for debate regarding this convention and its accordion-like history[75] did not trouble the Court in *Hall*. To the contrary, the majority took the use of two SDs as an unquestioned starting point in defining the kind of intellectual disability that precludes capital punishment. With that value in place, the case turned on a second convention — one pertaining to measurement error. As stated in Part I, the majority held that a cut score of $z = -2$ would be satisfactory for true scores but was constitutionally deficient for real scores (those with nonzero measurement error). Measurement error prevented a state from circumventing the need for a more comprehensive (and softer) clinical evaluation of offenders with IQ scores in the range of $z = -1.67$ to $z = -2$.

---

likely to give false confessions, are often poor witnesses, and are less able to give meaningful assistance to their counsel." *Id*. at 1993. Thus, an ideal test for intellectual disability would measure abilities with regard to deliberation, communication, and as well as the tendency to submit to authority.

Hall did not discuss the extent to which IQ scores (and the other parts of the current clinical nosology) match the constitutionally derived personal characteristics. It assumed that the measures of intellectual disability adopted for a variety of other purposes—including "education, access to social programs, and medical treatment plans," *Hall*, 134 S.Ct. at 1993 —also validly advance the penological and trial-process objectives identified in *Atkins*. The dissent thought that 70 was a constitutionally reasonable line of demarcation under *Atkins*—even if it no longer conformed precisely to clinical practice. Indeed, the dissent complained that the clinical definition was unstable, *id*. at 2006 (Alito, J., dissenting), and that the one-size fits all definition might not be as valid as one designed for capital-punishment purposes. *Id*. at 2006–07. In the dissent's view, "[p]ractical problems like these call for legislative judgments, not judicial resolution" (*id*. at 2007) based on the *Diagnostic and Statistical Manual* of the day. However, the divergence between the majority and the dissent is not really over *whether* to use IQ scores in ascertaining intellectual disability. It is over *how* states may use them.

[73] *See* Sadock & Sadock, *supra* note 61, at 2592 ("Given the Gaussian distribution of I.Q., as many people have an I.Q. between 70 and 74 as have an I.Q. between 0 and 69.").

[74] Of course, at some point, a proposed definition would be unreasonable. Situations like this constantly arise with vague predicates. Dominic Hyde, *Sorites Paradox*, in THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed. 2014), http://plato.stanford.edu/archives/win2014/entries/sorites-paradox/. They are woven into the very fabric of the law. *See* LEO KATZ, WHY THE LAW IS SO PERVERSE 157 (2011).

[75] *See supra* note 63+2.

In the next Part, I try to clarify what statistical theory has to say about the constitutionally tolerable location of the threshold for a conclusive presumption that the offender is not intellectually disabled. This Part considers, corrects, and builds on the Justices' treatment of the following statistics and ideas: (1) the meaning of a test-taker's "true score"; (2) estimates of test reliability; (3) estimates of the test's standard error of measurement; and (4) the confidence interval for the test-taker's true score.

III. True Scores and Single-measurement Error Within Classical Test Theory

A. First- and Second-order Questions

Why is $z = 1.67$ standard deviations below the mean — the figure that Court seemed to accept — the highest that the Constitution permits? Why not $z = 2.00$ as Florida wanted? One possible answer is that the 2.47% of the population whose scores lie between these two values includes too many people who are actually intellectually disabled and hence death-disqualified. Presumably, this is what the Court meant when it spoke of "an unacceptable risk that persons with intellectual disability will be executed."[76] but Justice Kennedy's opinion does not quantify this risk or the trade-off with the risk of not executing an offender who, under existing law, otherwise deserves to die. The only reason the Court offered to suspect that there are a substantial fraction of intellectually disabled capital offenders in this IQ range is that psychologists and psychiatrists have guidelines that use $z = -2$ as a benchmark (because scores this low are kind of rare) but then blur the boundary by referring to the possibility of measurement error in individual cases. For example, the American Association on Intellectual Development and Disabilities (AAIDD) states that a "significant limitation[] in intellectual functioning" exists when "[a]n IQ score . . . is approximately two standard deviations below the mean, considering the standard error of measurement for the specific assessment instruments' used and the instruments' strengths and limitations."[77]

The majority ran with this idea. As indicated in Part II, it adopted the conventional choice of $z = -2$ (corresponding to 70 points on the usual IQ-score scale), for the line that separates the intellectually able from the intellectually disabled but then argued that the "inherent error" in IQ tests means that to be reasonably sure of drawing the line at 70 (the figure selected because the number of people whose scores fall at or below 70 is not too large), it must be drawn at 75. Adjusting for measurement error in this manner represents imposing a second convention (relating to the uncertainty in individual measurements) on top of the first (establishing the departure from the average that suffices to permit a diagnosis of intellectual disability).

This two-part framework presupposes that the choice of $z = -2$ is at or near the edge of what is constitutionally permissible. If Florida could have drawn the first-order line substantially below

---

[76] *Hall*, 134 S. Ct. at 1990.

[77] AAIDD, INTELLECTUAL DISABILITY: DEFINITION, CLASSIFICATION, AND SYSTEMS OF SUPPORTS (11th ed. 2010) (as quoted in CHRIS HATTON, INTELLECTUAL DISABILITIES—CLASSIFICATION, EPIDEMIOLOGY AND CAUSES IN CLINICAL PSYCHOLOGY AND PEOPLE WITH INTELLECTUAL DISABILITIES 1, 4 (Eric Emerson et al. eds., 2d ed. 2012).

this value, then $z = -2$ already contains a tolerable margin for error. The failure to explain how the AAIDD-APA choice of an IQ score designed to prevent stigmatizing too many people as intellectually disabled captures the reasons expressed in *Atkins* for regarding mental retardation as a death-disqualifying condition exposes Justice Kennedy's opinion to the objection that "definitions of intellectual disability . . . are promulgated for use in making a variety of decisions that are quite different from the decision whether the imposition of a death sentence in a particular case would serve a valid penological end."[78]

Of course, direct attention to the validity of IQ scores for the desired purpose could lead to a higher rather than a lower score for precluding further inquiry into the existence of the disability.[79] My point is not that a state should be free to impose a stricter limit. It is simply to underscore that the *Hall* Court does not come to grips with the basic question in defining an IQ cut-score.[80]

Pretermitting that question, the Court pursued the second-order question of measurement error. Let us assume that $\mu_x - 2\sigma_x = 70$ is as high as one can go in demarcating the range of IQ scores that are so aberrant as to permit the diagnosis of intellectual disability. How does it follow that 75 (or some similar number) is the lower bound for scores that make it unnecessary to consider the adaptive functioning prong of the diagnosis? The Court invoked the "'standard error of measurement,' . . . often referred to by the abbreviation 'SEM.'"[81] The majority opinion then quoted the DSM-5, which states, "Individuals with intellectual disability have scores of approximately two standard deviations or more below the population mean, including a margin for measurement error (generally ±5 points). . . . [T]his involves a score of 65–75 (70 ± 5)."[82]

The dissent disputed the assertion that ±5 points is the "margin for measurement error." An exasperated Justice Alito complained that "there is no reason to assume a SEM of 5 points"[83] when "we know that the SEM for Hall's most recent IQ test was 2.16 — less than half of the Court's

---

[78] *Hall*, 134 S.Ct. at 2006 (Alito, J., dissenting). Justice Alito elaborated that"[i]n a death-penalty case, intellectual functioning is important because of its correlation with the ability to understand the gravity of the crime and the purpose of the penalty, as well as the ability to resist a momentary impulse or the influence of others. By contrast, in determining eligibility for social services, adaptive functioning may be much more important." *Id*. at 2006–07 (citations omitted).

[79] *See supra* note 72.

[80] Compared to the first-order question of the extent of the deficit in cognitive functioning required for the death-penalty exemption to serve the purposes identified in *Atkins*, the wobbliness in measured scores resembles a perturbation in the orbit of a planet. Because of Neptune's pull, Uranus was not quite where it should have been (as predicted from its gravitational interaction with only the sun), but astronomers had no trouble locating it before they discovered Neptune. The discrepancies were significant for a different reason. Calculations based on the small perturbations led to the discovery of Neptune. A. Pannekoek, *The Discovery of Neptune*, 3 CENTAURUS 126 (1953).

[81] *Hall*, 134 S. Ct. at 1995.

[82] APA, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS 37 (5th ed. 2013).

[83] *Hall*, 134 S. Ct. at 2010 (Alito, J., dissenting).

estimate of 5."[84] To evaluate this exchange, and to see whether *Hall* permits a state to specify a cutoff below 75, we must appreciate how the SEM is used to produce "a margin of error" and what alternatives are available. Unless one believes that anything that the APA and AAIDD do must be followed (even if there are statistically superior approaches that would satisfy the concerns articulated in *Atkins*), it becomes necessary to consider (1) what figure to use for the standard deviation of errors and (2) how many such standard errors are required to keep the risk of misclassification tolerable. We will see that Justice Kennedy could have avoided the ambiguities detected by Justice Alito by expressing the margin of error in terms of some fixed number of SEMs rather than the number 75. And looking more deeply, we will see that the Court's SEM is not the best standard error to use for an individual's score and that there are other reasonable ways to use the most appropriate statistic. A constitutional analysis ought to reflect *these* statistical facts as well as the popularity of the SEM.

B. True Scores and Measurement Error

Individual IQ scores are squishy. They are not perfectly reproducible. Measured scores fluctuate around "true scores," and to the extent that the observations erratically depart from the true value, their "reliability" is less than 1. In psychometrics, classical true-score theory defines an individual's unknown "true score" (which we can denote with the Greek letter tau, τ) as the average score that a given individual would achieve after taking the same test an infinite number of times (without learning anything from taking the test each time and without any other relevant changes). The difference between this test-taker's score x on a given test administration and the true score τ is the error (e) in that measurement (X). In these symbols, $x = \tau + e$. If the error is positive, the true score is less than the measured score. If it is negative, the true score is greater than the observed score. We know the observed score. We want to infer the true score.[85]

We can at least imagine many measurements of the same individual's IQ. Some of the errors in these measurements will be large, some small. These errors e will have some distribution with a particular shape that gives rise to a mean and a standard deviation. Classical true score theory posits that the errors e in the individual's repeated scores are normally distributed with mean 0 and an unknown standard deviation. The mean error ($\mu_e = 0$) indicates that the errors are centered on the true score, and the standard deviation (σ_e) tells how spread out they usually are. If we place true scores τ along the x-axis, the observed scores come from the normal distributions that could be drawn around each true score in Figure 2.

---

[84] *Id*. at 2011.

[85] The definition of "true score" as a long-term average for a test-taker on a particular test does not imply that the unobserved (latent) true score actually measures the construct that the test is designed to measure. John C. Willse, *Classical Test Theory*, *in* 1 ENCYCLOPEDIA OF RESEARCH DESIGN 149 (Neil J. Salkind ed. 2010). Validation of the test as a measure of "intelligence" or any other construct is a separate exercise. *Id*. at 152.

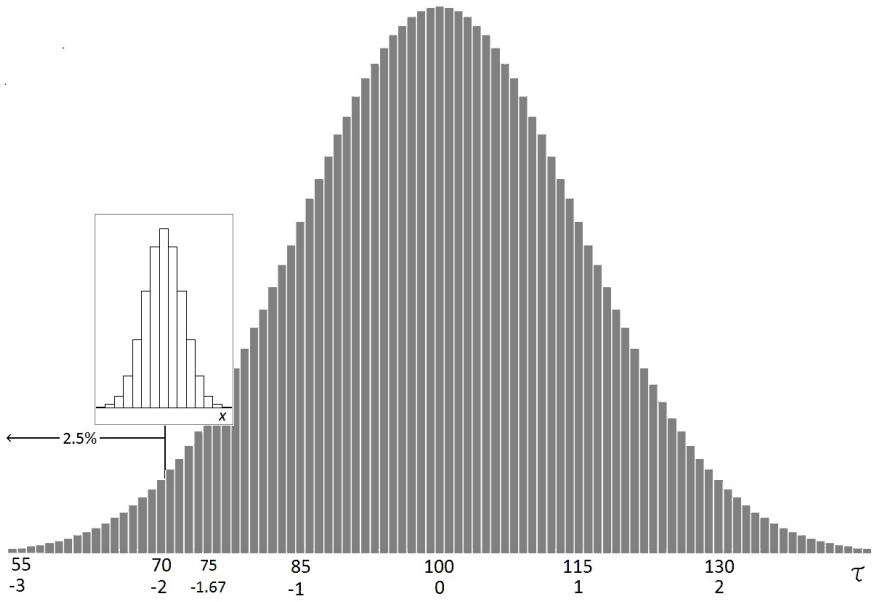

Figure 2. Theoretical Distribution of Observed Scores *x* Shown for One Individual with a True Score of $\tau = 70$ in a Population with Normally Distributed True Scores with Mean $\mu_\tau = 100$ and Standard Deviation $\sigma_\tau = 15$. The Individual's Observed Score Distribution (Indicated in the Smaller Histogram) Is Normal with Mean $\mu_x = \tau = 70$ and Standard Deviation $\sigma_e = 2.16$.

These normal distributions attached to the true scores must not be confused with the normal distribution that applies to the true IQs in a population. The error distributions for each possible true score apply to individuals at every true score level. They could be different even for test-takers who have the same true scores. Only one such distribution, for one test-taker, appears in Figure 2. Different individuals at each true-score level would have distributions with larger or smaller widths, since some people would be more consistent in their scores on repeated tests. The mean of each individual's error distribution is $\mu_e = 0$,[86] not $\mu_x = 100$, and its standard deviation σ_e for a given individual's error distribution is much smaller that the standard deviation of true scores across the entire population of heterogenous people. Smaller, but not zero. To assess the precision of the estimated true score τ of a given individual, we need to estimate σ_e for that individual.

C. Reliability and Standard Error of Measurement (SEM)

The Court fully recognized that IQ scores of different individuals are expected to fluctuate randomly around the true scores for these individuals. The majority put the point as follows:

[86] The mean of 0 makes the observed score *x* an unbiased estimator of the true score — sometimes, it will underestimate the true score; other times, it will overestimate τ; in the long run, these errors will average out to zero. It also is a maximum-likelihood estimator; no other value for τ has a higher probability of producing the measured value *x*. LLOYD ROSENBERG, STATISTICAL REASONING 18–23 (1971).

> Each IQ test has a "standard error of measurement" . . . often referred to by the abbreviation "SEM." A test's SEM is a statistical fact, a reflection of the inherent imprecision of the test itself. . . . An individual's IQ test score on any given exam may fluctuate for a variety of reasons. These include the test-taker's health; practice from earlier tests; the environment or location of the test; the examiner's demeanor; the subjective judgment involved in scoring certain questions on the exam; and simple lucky guessing.[87]

The SEM is an approximation of an average σ_e for all test-takers.[88] Instead of testing a single person many times to estimate σ_e, which is hardly feasible, test developers could test many people twice. Assuming that the test and retest scores are obtained within a short enough time period that each individual's true IQ itself has not changed[89] and that no individual improves with practice, the correlation between the test and the retest scores would indicate the reliability of the test. If the two scores were always identical, one could predict the second score perfectly from the first—the correlation would be 1. If they were completely unrelated, the first score would be of no value in predicting the second, and the correlation would be zero. Denoting the correlation in the test-retest scores for an entire population by the Greek letter rho (ρ), it can be shown that

$$\text{SEM} = \sigma_x \sqrt{(1 - \rho)}. \qquad (1)^{90}$$

If σ_x is 15, as the Court proposed, and if the reliability of the test is, say, $\rho = 0.82$,[91] then the SEM is $15 \sqrt{.18} = 6.36$.

In practice, test developers use less direct methods to estimate reliability. Roughly speaking, they split the test in half and compare the score on one half with that on the other half.[92] These two

---

[87] *Hall*, 134 S.Ct. at 1995.

[88] Won-Chan Lee et al., *Interval Estimation for True Raw and Scale Scores Under the Binomial Error Model*, 31 J. EDUC. & BEHAV. STAT. 261, 261 (2006) ("The traditional definition of SEM (i.e., same for all examinees) is sometimes called the overall SEM in the sense that it is an average SEM for all examinees in the population.").

[89] A study finding major changes is Sue Ramsden et al., *Verbal and Non-verbal Intelligence Changes in the Teenage Brain*, 479 NATURE 113 (2011).

[90] *See, e.g.,* MICHAEL FURR & VERNE R. BACHARACH, PSYCHOMETRICS: AN INTRODUCTION 119 (2d ed. 2014); David M. Lane, Measurement, *in* Online Statistics Education: A Multimedia Course of Study (David M. Lane ed.), http://onlinestatbook.com/2/research_design/measurement.html.

[91] *See* Marley W. Watkins & Lourdes G. Smith, *Long-Term Stability of the Wechsler Intelligence Scale for Children—Fourth Edition*, 25 PSYCHOL. ASSESSMENT 477, 480 (2013).

[92] *See* I. C. McManus, *The Misinterpretation of the Standard Error of Measurement in Medical Education: A Primer on the Problems, Pitfalls and Peculiarities of the Three Different Standard Errors of Measurement*, 34 MED. TEACHER 569, 571 (2012):

internal scores are analogous to test-retest scores. It is as if each person took two tests at once. If the internal reliability is $\rho = 0.98$,[93] then the SEM is $15 \sqrt{.02} = 2.12$.

As this example indicates, the Court's understanding of the overall SEM as an inherent property of the test is not quite correct.[94] The SEM can vary with the group used to estimate test-retest reliability; moreover, other measures of reliability are also employed (to account for measurement error induced by different forms of the same test or scoring by different raters).[95] As Justice Alito wrote, "there is not a single, uniform SEM across IQ tests or even across test-takers. Rather, 'the [SEM] varies by test, subgroup, and age group.'"[96]

This complication, however, does not preclude estimates of σ_e for different test-takers. As is typical in applied statistics, there will be arguments about the application of the general concepts

---

> Conceptually, test reliability is typically considered as how a large group of candidates would perform if the identical assessments were taken on two separate occasions (the test–retest correlation). Most high-stakes assessments, though, are taken but once, and instead there are several ways in which reliability can be calculated from the internal structure of the test. The split-half correlation is analogous to test–retest correlation, comparing the performance of candidates on, say, odd-numbered items and even-numbered items. Cronbach's alpha which statisticians prefer, is a generalisation of the split-half correlation across all possible ways of dividing a test.

[93] The WAIS-IV test has a reliability of .97–.98 based on internal consistency, although "[t]hese should be considered best-case estimates because they do not consider other major sources of error such as long-term temporal stability, administration errors, or scoring errors." Gary L. Carnivez, *Review of Wechsler Adult Intelligence Scale—Fourth Edition*, *in* THE EIGHTEENTH MENTAL MEASUREMENTS YEARBOOK (Robert A. Spies et al. eds. 2010). By some internal measures, the reliability of full IQ scores on the Stanford-Binet test is between 0.91 and 0.98. Sherry K. Bain & Jessica D. Allin, *Book Review: Stanford-Binet Intelligence Scales: Fifth Edition*, 23 J. PSYCHOEDUCATIONAL ASSESSMENT 87, 90 (2005).

[94] *See, e.g.,* Lane, *supra* note 90 ("If a test were given in two populations for which the variance of the true scores differed, the reliability of the test would be higher in the population with the higher true-score variance. Therefore, reliability is not a property of a test per se . . . .").

[95] *See, e.g.,* McManus, *supra* note 92, at 571:

> Conceptually, test reliability is typically considered as how a large group of candidates would perform if the identical assessments were taken on two separate occasions (the test–retest correlation). Most high-stakes assessments, though, are taken but once, and instead there are several ways in which reliability can be calculated from the internal structure of the test. The split-half correlation is analogous to test–retest correlation, comparing the performance of candidates on, say, odd-numbered items and even-numbered items.3 Cronbach's alpha which statisticians prefer, is a generalisation of the split-half correlation across all possible ways of dividing a test.

[96] *Hall*, 134 S.Ct. at 2009 (Alito, J., dissenting) (quoting User's Guide To Accompany AAIDD 11th ed.: Definition, Classification, and Systems of Supports 22 (2012)).

and group statistics to individuals,[97] but let us assume that one has a good estimate of reliability for the kinds of random measurement error that are of concern for the test and the defendant who has taken it. If the reliability, and hence the overall SEM, accounts for all the important sources of departures from the individual's true score, what can we say about the individual's true score? The majority thought that the answer lies in using the overall SEM to form a confidence interval around the observed score.[98]

D. Confidence Intervals from the SEM (SEM-IS)

Justice Kennedy asserted that "[t]he SEM allows clinicians to calculate a range within which one may say an individual's true IQ score lies."[99] He added that "SEM is a unit of measurement: 1 SEM equates to a confidence of 68% that the measured score falls within a given score range, while 2 SEM provides a 95% confidence level that the measured score is within a broader range."[100] This exposition is not ideal. Clinicians do not use the SEM to decide whether "*the measured score* falls within a given score range."[101] The measured score is the IQ score x, and one can say with 100% confidence whether this score falls within any interval one likes. Moreover, as explained below, "confidence" is a term of art — 95% "confidence" does not have the simple meaning that a lay reader would ascribe to it (and that the Court may have thought it does).

---

[97] *See generally* David L. Faigman et al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417 (2014); Nicholas Scurich & Richard S John, *A Bayesian Approach to the Group Versus Individual Prediction Controversy in Actuarial Risk Assessment*, 36 LAW & HUM. BEHAV. 237 (2012).

[98] The dissent suggested that an SEM tailored to the individual should be used. *Hall*, 134 S.Ct. at 2010 (Alito, J., dissenting) (referring to "the SEM for a particular test and a particular test-taker").

[99] *Hall*, 134 S.Ct. at 1995. Going into greater detail, an amicus brief from the American Psychological Association and other professional organizations repeatedly cited by the Court clumsily stated that "[t]he SEM . . . is . . . used to calculate confidence intervals. Thus, a full scale IQ 'score of 70 is most accurately understood not as a precise score but as a range of confidence with parameters of at least one standard error of measurement.'" Brief for American Psychological Association et al. as Amici Curiae in Support of Petitioner, Hall v. Florida, No. 12-10882, at 23 (quoting AAIDD Manual at 24). This phrasing is statistically inept. Certainly, a score of 70 is not precise—the true score could be higher or lower—but how much higher or lower is not "a range of confidence" accompanied by "parameters." No "parameters" of a statistical model accompany an interval estimate of $\pm k$ SEM around the measured score. The confidence coefficient is a single number that defines k. There is no fundamental reason to specify at "at least one SEM" ($k \geq 1$) as the relevant range.

[100] *Id*. (quoting Brief for American Psychological Association et al., *supra* note 99). Calling the SEM a "unit of measurement" is confusing — an IQ test's unit of measurement is an IQ point. Of course, the fact that IQ scores can be transformed into other units such as numbers of SEMs from the mean permits the transformed score to be used instead of the reported IQ scores. In that sense, any transformation that can be inverted could be said to define a new unit of measurement.

[101] *Id*. (emphasis added).

In an act of statistical one-upmanship, Justice Alito supplied examples of how to compute confidence intervals with the SEM. According to the dissent,

> If a test-taker scores a 72 on an IQ test with a SEM of 2, the 66% confidence interval is the range of 70 to 74 (72 ± 2). In this situation, there is approximately a 66% chance that the test-taker's "true" IQ is between 70 and 74; roughly a 17% chance that it is above 74; and roughly a 17% chance that it is 70 or below. Thus, there is about an 83% chance that the score is above 70.
>
> Similarly, using two SEMs, we can build a 95% confidence interval. [L]et us hypothesize a case in which the defendant's obtained score is 74. With the same SEM of 2 as in the prior example, there would be a 95% chance that the true score is between 70 and 78 (74 ± 4); roughly a 2.5% chance that the score is above 78; and about a 2.5% chance that the score is 70 or below. The probability of a true score above 70 would be roughly 97.5%.[102]

The arithmetic is mostly correct.[103] The semantics are not. It is common to form a confidence interval (CI) by adding and subtracting some number of SEMs to the observed score x.[104] And to increase "confidence," one need only add and subtract a larger number of SEMs.[105] To this extent, the opinion is correct. But "confidence" is an abstruse and technical term.[106] It is not, as Justice Alito proposed, the probability or "chance that the test-taker's 'true' IQ falls within this range."[107] In

---

[102] *Hall*, 134 S.Ct. at 2010 (Alito, J., dissenting).

[103] A trivial but obvious correction (to readers familiar with the normal distribution) is that an interval of ±1 standard deviation covers 68%, not 66%, of the area under the curve.

[104] FURR & BACHARACH, *supra* note 90, at 169–70.

[105] *Hall*, 134 S.Ct. at 2010 (Alito, J., dissenting) ("the greater the degree of confidence demanded, the greater the range of scores that will fall within the confidence interval").

[106] *E.g.*, MORRIS H. DEGROOT & MARK J. SCHERVISH, PROBABILITY AND STATISTICS 412 (3d ed. 2002) ("Because of the distinction between confidence and probability, the meaning and relevance of confidence intervals in statistical practice is a somewhat controversial topic."); David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 247 (Federal Judicial Center & National Research Council Committee on the Development of the Third Edition of the Reference Manual on Scientific Evidence eds., 3d ed. 2011) ("the confidence level does not give the probability that the unknown parameter lies within the confidence interval. . . . According to the frequentist theory of statistics, probability statements cannot be made about population characteristics: Probability statements apply to the behavior of samples. That is why the different term 'confidence' is used.").

[107] *E.g.*, DEGROOT & SCHERVISH, *supra* note 106, at 412 ("It should be emphasized that it is *not* correct to state that θ [the parameter being estimated] lies in the interval (a, b) with *probability* γ [the confidence coefficient]") (emphasis in original); KAYE ET AL., THE NEW WIGMORE ON EVIDENCE: EXPERT EVIDENCE § 12.6.4, at 546–47 (2d ed. 2011) ("confidence of 95 percent does not necessarily mean that the interval estimate has a 95 percent probability of being correct."); Kaye & Freedman, *supra* note 106, at 247.

classical true score theory, the true score is a fixed number, not a random variable with probabilities distributed across some range of possible values. In Justice Alito's second example, 74 ± 4 is one of many possible 95% CIs. If we retested the same person, the second score might be 76, giving rise to a new CI of 76 ± 4. A third test might yield a CI of 80 ± 4. As the number of intervals constructed from the formula CI = x ± 1.96 SEM grows infinitely large, 95% of them will include the true score τ, whether that score is between 70 and 78, whether it is between 72 and 80, whether it is between 76 and 80, or whether it is inside any other conceivable interval. Thus, the CI is useful for indicating the precision of an estimate,[108] but ascertaining the probability that the true score is within a given region requires other methods.[109]

The conflation of "confidence" with "probability" leads the dissent to accuse the majority of misallocating the burden of persuasion. Justice Alito wrote that

> As Hall concedes, the Eighth Amendment permits States to assign to a defendant the burden of establishing intellectual disability by at least a preponderance of the evidence. . . . In other words, a defendant can be required to prove that the probability of a 70 or sub–70 IQ is greater than 50%. Under the Court's approach, by contrast, a defendant could prove significantly subaverage intellectual functioning by showing simply that the probability of a "true" IQ of 70 or below is as little as 17% (under a one-SEM rule) or 2.5% (under a two-SEM rule). This totally transforms the allocation and nature of the burden of proof.
>
> [I]t would be simple enough to devise a 51% confidence interval—or a 99% confidence interval for that matter. There is therefore no excuse for mechanically imposing standards that are unhinged from legal logic and that override valid state laws establishing burdens of proof. The appropriate confidence level is ultimately a judgment best left to legislatures, and their judgment has been that a defendant must establish that it is more likely than not that he is intellectually disabled. I would defer to that determination.[110]

To continue where Justice Alito left off in his construction of CIs, the 51% CI centered on an observed score extends ±0.69 standard deviations around that score. For the postulated σ_e = SEM = 2, a score of x = 72 produces a 51% CI with a lower bound of 71;[111] a score of x = 71 yields a lower bound of 70.[112] According to Justice Alito's analysis, therefore, a state that defined "significantly subaverage intellectual functioning" as a true IQ score of 70 could exclude all offenders with observed scores (on that IQ test) of 72 or more from further consideration on the theory that it is probably the case that these offenders' true scores exceed 70.

---

[108] Kaye & Freedman, *supra* note 106, at 247–48.

[109] *See* Kaye, *supra* note 10; *infra* Part IV.B.

[110] *Hall*, 134 S.Ct. at 2011 (Alito, J., dissenting).

[111] The exact number is 70.6.

[112] The number on the continuous scale is 69.6.

There are two problems with this reasoning. First, Justice Alito stops short of the logical conclusion to the argument. On the dissent's understanding of "confidence" and the burden of persuasion, there is no reason to compute a confidence interval. After all, we are only concerned with errors in the lower tail of the curve. (By definition, a defendant whose true IQ score is higher than the interval estimate for any observed score of at least 70 is not intellectually disabled.) The normal curve is symmetric, so exactly 50% of the area under the curve lies to the left of the measured score. Look back at Figure 2. When the distribution is centered on 70 itself, 50% of the error distribution dips below 70. When it is centered on 71, less than 50% of the error distribution is below 70. The same is true of the error distribution around an observed score of 72, 73, and so on. Consequently, *no* confidence interval at all need be used. Every measured score greater than 70 has the majority of its measured error in the region above 70. No computation is required.

The second and more fundamental problem is that "confidence" is not a probability statement about a fact subject to the burden of persuasion. As I show in a more detailed essay on "confidence" and the burden of persuasion, depriving the state of the opportunity to fix the line at either 70 (considering only the lower tail of the error distribution) or 72 (considering both tails) does not necessarily conflict with the state's general use of a 51% burden of persuasion; and it is not at all obvious that the more-probable-than-not standard should apply in this context.[113] More defensible procedures for establishing a cut-score for measured IQ scores are available. The next section presents one of them.

E. SEM-adjusted-maximum Score (SEM-AM)

Justice Alito defended the 70-point cutoff as much simpler to apply than the majority's incorporate-a-margin-of-error approach. He remarked that

> it is unclear to me whether the Court concludes that a defendant is constitutionally entitled to introduce non-test evidence of intellectual disability (1) whenever his score is 75 or lower, on the mistaken understanding that the SEM for most tests is 5; (2) when the 66% confidence interval (using one SEM) includes a score of 70; or (3) when the 95% confidence interval (using two SEMs) includes a score of 70.[114]

Unfortunately, Justice Kennedy did not clarify which of these possibilities the Court had in mind, but it is doubtful that the references to the 70–75 range and to Hall's concession that a state could choose 75 as an absolute cutoff are integral parts of its holding. The disability-testing guidelines now tend to be framed in terms of standard errors rather than specific scores.[115] The Court simply

---

[113] Kaye, *supra* note 10.

[114] *Hall*, 134 S.Ct. at 2011 (Alito, J., dissenting).

[115] For example, the American Association on Intellectual and Developmental Disabilities refers to the requisite indication of a deficiency in "[i]ntellectual functioning" as "[a]n IQ score that is approximately two standard deviations below the mean, considering the standard error of measurement for the specific assessment instruments' used and the instruments' strengths and limitations." AAIDD, *supra* note 60 (as

"agree[d] with the medical experts that when a defendant's IQ test score falls within the test's acknowledged and inherent margin of error [around the statutorily prescribed cutoff for eligibility], the defendant must be able to present additional evidence of intellectual disability, including testimony regarding adaptive deficits."[116] It is not that hard to use a rule that sets the maximum IQ score for a finding of intellectual disability at the legislative maximum for the true score (such as 70) plus 1.96 SEMs for the test in question.[117] For an IQ test with an SEM of 2.16, this maximum would be $70 + 1.96 \times 2.16 = 74$. A score on this test of 75 or above thus would make an offender ineligible for the *Atkins* exemption.

We can call such a rule an SEM-adjusted-maximum-score rule (SEM-AM, for short) because it takes the legislatively determined true score and boosts it some number of SEMs to fix the maximum IQ score from a single test that can place an offender (subject to other evidence) in the intellectually disabled range. In symbols, the adjusted maximum is $x_{max} = \tau_{leg} + k$ SEM, where τ_{leg} is the legislative choice for the largest true score consistent with intellectual disability and k is the number of SEMs for the desired "confidence" (68% for k =1, and 95% for k = 1.96). If SEM were

---

quoted in HATTON, *supra* note 60, at 4 ("Individuals with intellectual disability have scores of approximately two standard deviations or more below the population mean, including a margin for measurement error (generally +5 points). On tests with a standard deviation of 15 and a mean of 100, this involves a score of 65–75 (70 ± 5)."); COMMITTEE ON DISABILITY DETERMINATION FOR MENTAL RETARDATION, *supra note* 63, at 115 ("no matter how great the discrepancy between relevant subscales, individuals with total test scores greater than 75 should not be diagnosed as having mental retardation.").

[116] *Hall*, 134 S.Ct. at 2001.

[117] Although it is not obvious how much of the theory of true-score estimation the majority really understood, the dissent hardly seems justified in attributing to the Court "the mistaken understanding that the SEM for most tests is 5." Justice Kennedy clearly stated that "the average SEM for the WAIS–IV is 2.16 IQ test points and the average SEM for the Stanford–Binet 5 is 2.30 IQ test points . . . ." *Hall*, 134 S.Ct. at 1995. Justice Alito seems convinced that using more than one SEM for the necessary "margin of error," as the majority plainly did, somehow contravenes professional standards accepted in *Atkins*. But there is little basis for that view. According to the dissent:

> [T]he Court misreads the authorities on which it relies to establish this cutoff IQ score of 75. It is true that certain professional organizations have advocated a cutoff of 75 and that *Atkins* cited those organizations' cutoff. See ante, at 12, 20. But the Court overlooks a critical fact: Those organizations endorsed a 75 IQ cutoff based on their express understanding that " one standard error of measurement [SEM]" is "three to five points for well-standardized" IQ tests. AAMR, Mental Retardation 37 (9th ed.1992) (hereinafter AAMR 9th ed.); *Atkins*, 536 U.S., 309, n. 5 (citing AAMR 9th ed.; 2 Kaplan & Sadock's 2592 (B. Sadock & V. Sadock eds., 7th ed. 2000)); see also AAMR 10th ed. 57; AAIDD 11th ed. 36. In other words, the number 75 was relevant only to the extent that a single SEM was "estimated" to be as high as 5 points. AAMR 9th ed. 37.

Id. at 2010-11. Given the long-established tradition of the 0.05 level of statistical significance and its cognate 95% CI in psychology, it would be surprising if these authorities favored a "margin of error" of 1 SEM in the 1990s.

equal to the standard error σ_e for every defendant, one could say that the SEM-AM rule for 95% confidence would keep only about 2.5% of offenders with true IQ scores at τ_{leg} from producing further evidence of their disability.[118]

In this way, a state can pick a single number x_{max}, as Justice Alito wanted. Moreover, this simple procedure does the same thing as the SEM-IS rule the dissent deprecated as unduly complicated. In statistical jargon, the SEM-AM rule is simply a null hypothesis test that tells us whether, at the 0.025 level for a false alarm, an observed score is larger than 70.[119] The outcomes of this test for statistical significance are identical to the decisions that follow from using two-sided 95% CIs in the SEM-IS procedure. That is, if we decide that a true score is above 70 if and only if the 95% CI for the measured score x is entirely above 70, we will reject the claim of a true IQ of 70 or less in just those cases in which the measured IQ x exceeds 70 + 1.96 SEM.[120] Whether we decide according to the SEM-IS or the SEM-AM rule makes no difference.[121]

Two caveats on these rules are in order. First, we cannot specify the expected proportion of defendants who are falsely classified as ineligible, for that depends on how many convicted capital offenders have true scores of 70, 69, 68, 67, and so on. For example, if we were to assume that 30% have true IQs of $\tau = 70$, that 15% have $\tau = 69$, that 8% have $\tau = 68$, and that 4% have $\tau = 67$, then the expected error rate would be (30% × .025) + (15% × .0077) + (8% × .0020) + (4% × .0004) + ... = 0.88%.[122]

Second, as noted earlier, the SEM is a kind of average error across all IQ scores. As an estimator of σ_e, it works best for estimating the variation in true scores that are near the mean for all of these test-takers. But the only offenders who can seek the exemption from capital punishment are

---

[118] The expected error rate of 2.5% applies to offenders with true IQ scores of exactly τ_{leg}. It ignores errors with respect to offenders with smaller true scores. Some proportion of offenders with true scores below τ_{leg} also will have a single IQ test score larger than $x_{max} = \tau_{leg} + k$ SEM. They too will be precluded from presenting evidence of adaptive functioning even though they meet the true-score criterion for intellectual disability. For example, when the legislative choice of the true score is 70 and the SEM is 2.16 (yielding x_{max} = 74.23), 0.77% of defendants with true scores of 69 will have a single observed score greater than x_{max}. Likewise, 0.20% of defendants with true scores of 68 will be falsely deemed ineligible. For a true score of 67, we are down to a 0.04% conditional error rate. As these numbers suggest, we can say that 0.025 is an upper bound on the probability that the hypothesis test will misclassify any individual whose IQ is truly τ_{leg} or less as having a true score $\tau > \tau_{leg}$. The probability of misclassification given an observed score $x > \tau_{leg}$ is another story. *See infra* note 122

[119] For definitions or explanations of hypothesis testing, see, for example, ROSENBERG, *supra* note 86, at 34–38; KAYE ET AL., *supra* note 107.

[120] For a graphical explanation, see Kaye, *supra* note 10.

[121] For a general proof that "a coefficient γ confidence set . . . can be thought of as a set of null hypotheses that would be accepted at significance level $1 - \gamma$," see DEGROOT & SCHERVISH, *supra* note 106, at 457. In this sense, significance tests underlie confidence intervals. D. R. COX, PRINCIPLES OF STATISTICAL INFERENCE 40 (2006).

[122] The factors of 0.025, 0.0077, 0.0020, and 0.0004 are the areas in the tail of normal curve up to 70, 71, 72, and 73, respectively. Part IV.B makes more use of the frequency distribution of true scores.

in the vicinity of two or more standard deviations from the population mean (if that is the targeted true score τ_{leg}). The most appropriate SEM for an SEM-AM rule might come from reliability statistics for a group of low scoring test-takers in the region of τ_{leg}.

F. Confidence Intervals from the Standard Error of Estimate (SEE-IS)

Let us return to the Court's main idea of constructing an interval estimate based on the observed score and seeing whether it goes low enough to be considered in the substantially subaverage zone necessary for a diagnosis of disability. The SEM-interval-score rule (SEM-IS) that we previously discussed used intervals of $x \pm k$ SEM. But psychometricians have long recognized that a different standard error is more suitable for inferring true scores from observed ones.[123] This "standard error of estimate" or SEE[124] indicates the variation in the true scores of different test-takers *who have the same measured score*. In other words, it is a conditional standard error of measurement rather than the broader, overall SEM. It pertains to the people who score the same as a given individual j. All of them have the one observed score x_j, but not all of them have the same true score τ_j. Some have τ above x_j, some have τ below x_j, and some have $\tau = x_j$. Assuming normality and equal variance of the error term for all test-takers, the conditional distribution of τ for the x_j-score cohort is normal with standard deviation

$$\text{SEE} = \sigma_x \sqrt{\rho}(1 - \rho) = \text{SEM} \sqrt{\rho}. \qquad (2)^{125}$$

Applying this formula to the previous examples of SEMs of 6.36 and 2.12 gives SEEs of 5.76 and 2.10. Conditioning on the observed score has reduced the standard error.

Conditioning has a second consequence. It shifts the point on which the interval sits toward the mean. Instead of using the observed score x as the point estimate, we use $\rho x + (1 - \rho)\mu_x$.[126] This adjustment reflects the common, but often unrecognized phenomenon of regression to the mean.[127]

---

[123] *E.g.*, Frank J. Dudek, *The Continuing Misinterpretation of the Standard Error of Measurement*, 86 PSYCHOL. BULL. 335 (1979); J. C. NUNNALLY, PSYCHOMETRIC THEORY 218 (1978); ELAZAR J. PEDHAZUR & LIORA PEDHAZUR SCHMELKIN, MEASUREMENT, DESIGN, AND ANALYSIS: AN INTEGRATED APPROACH 111–12 (1991).

[124] Richard A. Charter & Leonard S. Feldt, *Confidence Intervals for True Scores: Is There a Correct Approach?*, 19 J. PSYCHOEDUCATIONAL ASSESSMENT 350, 354–55 (2001); R. FURR & BACHARACH, *supra* note 90, at 171; McManus, *supra* note 92, at, 572 (using the abbreviation "SEest"); Richard B. McHugh, *The Interval Estimation of a True Score*, 54 PSYCHOL. BULL. 73, 73 n.1 (1957) (not using an abbreviation for "standard error of estimate").

[125] *See, e.g.,* Charter & Feldt, *supra* note 124, at 354; Dudek, *supra* note 27, at 335.

[126] *Id*.

[127] McManus, *supra* note 92, at 573. *See also* David H. Kaye, *The Disappearance that Wasn't? "Random Variation" in the Number of Women Supreme Court Clerks*, 48 JURIMETRICS J. 457 (2008) (noting that the phenomenon as a possible explanation for a highly publicized shortfall in women hired as clerks one year at the Supreme Court).

To the extent that chance affects test scores — the central concern of the *Hall* Court — some people will get a bonus over what they could earn based on their ability alone, and some will be penalized. Thus, chance will pull or push some people into the two extreme ends of the distribution of scores. On average, the truly high-ability people will stay high and the truly low-ability ones will stay low, but the moderate-ability people who happened to score very high or low on the test the first time around will drift back to the middle of the pack in the scenario of infinite retesting that produces a true score.[128]

As an example of applying these regression-based formulas for estimating the range of true scores that people with a given observed score have, consider an observed score of $x = 75$ on a test that has a reliability of $\rho = 0.97$. In the SEM-IS approach, the 95% CI would be $75 \pm 1.96 \times 15 \sqrt{(1 - .97)} = 70$ to 80. A defendant could be said to fall into the potentially disabled zone of 70 or below. In the regression-based SEE-IS approach, the 95% CI is $75.75 \pm 1.96 \times 15 \sqrt{.97(1 - .97)} = 71$ to 81. The net effect of narrowing and raising the interval in this example is that offenders with single measured IQ scores of 75 on this test no longer qualify for a more probing inquiry into their possible intellectual disability.

This procedure for correcting IQ scores for regression to the mean is related to far more general statistical ideas about shrinkage estimators.[129] These estimators[130] compromise the properties of more traditional estimators "(maximum likelihood, minimum variance unbiased, least squares,

---

[128] *See* David M. Lane, *Regression Toward the Mean*, ONLINE STATISTICS EDUCATION: A MULTIMEDIA COURSE, http://onlinestatbook.com/2/regression/regression_toward_mean.html.

[129] G. K. Robinson, *That BLUP is a Good Thing: The Estimation of Random Effects*, 6 STAT. SCI. 15 (1991), presents as "far from new" the following example:

> Suppose that true intelligence quotient (IQ) is normally distributed with mean 100 and standard deviation 15. Two tests are available. Both tests give scores that are normally distributed with mean the true IQ. The first test score has standard deviation 10 given true IQ, while the second test score has standard deviation 5. A person scoring 130 on the first test would be estimated to have a true IQ of 120.8 and a person scoring 130 on the second test would be estimated to have a true IQ of 127. Features of these estimates worth noting are as follows.
> - They are shrunk towards the overall mean (100) from the data.
> - The amount of shrinkage is greater when the data point is less informative.
> - They are biased given true IQ. This is obvious since the raw scores are unbiased and the estimates are nontrivial linear functions of the raw scores.
> - They have zero average bias when averaged over the distribution of possible true IQs.
> - The expected value of true IQ given the data is equal to the BLUP [best linear unbiased prediction] estimate of IQ . . . .

Id. at 22.

[130] Different types of shrinkage estimators are "ordinary shrinkage, preliminary test (shrinkage), Stein-type, ridge regression, empirical Bayes. estimators, etc." Hermanus H. Lemmer, *Shrinkage Estimators*, *in* ENCYCLOPEDIA OF STATISTICAL SCIENCE (Samuel Kotz & Campbell B. Read eds., 2d ed. 2006).

etc.)" to achieve better precision or other desirable qualities.[131] They have been the subject of considerable study[132] and application in many fields.[133]

The SEE-IS approach of using a conditional standard error permits statements about the frequency with which the true score would fall into the confidence interval. For example, one can legitimately report that "[f]or people who have a score like yours, we find that X% of them truly fall somewhere between A and B."[134] Following Justice Alito's guide to CIs, a defendant with a score of 71 on our test with a reliability of .98 could argue that about 68% of observed scores of 71 correspond to true scores in the 70–74 range; hence, 16% of them have true scores of 70 or below. Then, shifting to the majority's reasoning, the defendant could urge that depriving 16% of people who might otherwise be put to death of the opportunity to bring forth proof of an intellectual disability "creates an unacceptable risk that persons with intellectual disability will be executed, and thus is unconstitutional."[135]

IV. Other Statistical Issues in and Outside of *Hall*

The discussion up to this point does not exhaust the number of procedures for constructing confidence intervals that statisticians have devised.[136] *Hall* should be read as directing experts to handle measurement error in IQ scores in the most professionally responsible manner rather than instantiating any one of the narrow alternatives listed in the dissenting opinion. This would leave room for combining multiple scores and for the consideration of Bayesian procedures that assign probabilities to all possible true scores. This Part briefly discusses these two matters.

A. Multiple Scores

Justice Alito complained that "the Court entirely ignores [the fact] that Florida . . . takes into

---

[131] *Id*. ("to minimize (maximize) some desirable criterion function (mean square error, quadratic risk, bias, etc.))."

[132] *E.g.*, R. W. Farebrother, *A Class of Shrinkage Estimators*, 40 J. ROYAL STAT. SOC'Y. SERIES B 47 (1978); Dominique Fourdrinier & Martin T. Wells, *On Improved Loss Estimation for Shrinkage Estimators*, 27 STAT. SCI. 61 (2012).

[133] *E.g.*, FRANK HARRELL, REGRESSION MODELING STRATEGIES: WITH APPLICATIONS TO LINEAR MODELS, LOGISTIC AND ORDINAL REGRESSION, AND SURVIVAL ANALYSIS 75 (2d ed. 2015); G. S. Maddala et al., *A Comparative Study of Different Shrinkage Estimators for Panel Data Models*, 2 ANNALS ECON. & FINANCE 1 (2001); Yi-Hau Chen et al., *Shrinkage Estimators for Robust and Efficient Inference in Haplotype-Based Case-Control Studies*, 104 J. AM. STAT. ASS'N 220 (2009).

[134] Richard A. Charter & Leonard S. Feldt, *The Importance of Reliability as it Relates to True Score Confidence Intervals*, 35 MEASUREMENT & EVALUATION IN COUNSELING & DEVELOPMENT 104, 107 (2002).

[135] *Hall*, 134 S.Ct. at 1990.

[136] *See, e.g.,* Lee et al., *supra* note 88.

account the inevitable risk of testing error by permitting defendants to introduce multiple scores."[137] Indeed, the trial court in *Hall* heard testimony about at least four IQ test scores for Hall — 71, 72, 73, and 80[138] — and it appears that Hall had taken as many as seven IQ tests over a 40-year period.[139]

Although the majority did not "entirely ignore" multiple scores, Justice Kennedy's treatment of them is difficult to understand, statistically as well as legally. He wrote that "[e]ven when a person has taken multiple tests, each separate score must be assessed using the SEM,"[140] and he seemed to be referring to the single-score SEM. If this is what "the SEM" means, the dictum cannot be right. Uncertainty about the true score declines as more measurements are made.[141] In technical terms, because the SEM for a single score is greater than the standard error of the average of several scores, using the single-score SEM as a measure of the probable error in the average score would be a mistake. An SEM-based score interval for the average can lie above a figure like 70 even though at least one — or even all — of an offender's separate intervals dip below 70. Assuming that Hall's IQ scores are independent, as the dissent implied,[142] the 95% CI that applies to his average score is 73 to 75.[143] Having accounted for the measurement error in the average of the IQ scores, the state should be able to enforce a 70-or-less rule for a true score, barring Hall from producing evidence of his clear deficits in adaptive functioning..

One might argue that the computation of the interval for Hall's average is oversimplified, for it assumes fully independent scores. Indeed, the majority was worried that "the analysis of multiple

---

[137] *Hall*, 134 S.Ct. at 2007; *see also id*. at 2008 ("We have been presented with no solid evidence that the longstanding reliance on multiple IQ test scores as a measure of intellectual functioning is so unreasonable or outside the ordinary as to be unconstitutional.").

[138] *Id*. at 2007 n.9. It declined to consider a score of 69 on another test because the psychologist who administered and scored the test was dead and Hall's counsel had violated an order "to provide the State with the [underlying] testing materials and raw data." Brief for Respondent 19 n.11.

[139] Amended Order, Florida v. Hall, No. 1978-CF-0052, at ¶¶ 16, 20 (Fla. Cir. Ct. June 8, 2010) (Joint App. at 105, 108).

[140] *Hall*, 134 S.Ct. at 1995.

[141] Only if successive scores were completely dependent on one another would the additional scores fail to provide information as to the true score.

[142] *Hall*, 134 S.Ct. at 2007 n. 9 (Alito, J., dissenting) ("Hall does not allege that any potential "practice effect" skewed his scores.").

[143] The mean of Hall's IQ scores of 71, 72, 73, and 80 is 74. The *Hall* Court's 95%-confidence SEM-IS for a single score of 74 for an IQ test with SEM 2.16 goes from 70 to 78. Because it does not lie entirely above 70, Hall is eligible for the disability exemption—the chance that he is disabled (assuming, as the record strongly indicates, that he has deficits in adaptive functioning), is too large under the 95%-SEM-IS rule to permit the state to kill him. However, if the four scores are independent and if each test has the same SEM of 2.16, then the standard error *of the mean* is less than 2.16. It is 1/4 of the square root of the sum of the squared SEMs for the four tests, which equals 0.73. The 95% confidence interval for the mean is therefore $74 \pm 1.96 \times 0.73$, which goes from 73 to 75. This interval estimate puts Hall outside the IQ range that permits the diagnosis of intellectual disability.

IQ scores jointly is a complicated endeavor."[144] But this argument is a makeweight. Does the majority really believe that the Constitution requires the state to treat a defendant with ten successive, carefully administered IQ scores of 71 the same as a defendant with but a single score of 71?[145] Not only is this rule intuitively implausible, but analysis of multiple scores is hardly beyond the ken of statisticians.[146] The chapter in a psychology handbook[147] cited in Justice Kennedy's majority opinion for the view that the problem is unduly complicated[148] actually states that the solution "is not nearly so complicated as it might seem at first glance."[149] It shows how to combine scores with pencil and paper or a spreadsheet.[150]

Finally, Justice Kennedy added that "because the test itself may be flawed, or administered in a consistently flawed manner, multiple examinations may result in repeated similar scores, so that even a consistent score is not conclusive evidence of intellectual functioning."[151] Certainly, all IQ tests are imperfect in many ways. In addition to being unable to eliminate random error, they may be culturally biased or "biased in favor of neurotypical individuals."[152] Or, a defendant may be

---

[144] *Hall*, 134 S.Ct. at 1995.

[145] In remanding the case, the Court emphasized that "Freddie Lee Hall may or may not be intellectually disabled, but the law requires that he have the opportunity to present evidence of his intellectual disability, including deficits in adaptive functioning over his lifetime." *Hall*, 134 S. Ct. at 2001. It would have been more appropriate to remand for a determination of (1) whether the measurement error associated with all of Hall's measured scores (taken together) was such his true score very probably exceeded 70, and if not, (2) whether the scores, combined with the evidence on adaptive functioning, established an intellectual disability.

[146] Courts that are not advised of the appropriate statistical analysis may be tempted to reject a claim of intellectual disability if any score is two SEMs above 70. *E.g.*, Williams v. Stephens, 761 F.3d 561, 573 (5th Cir. 2014) ("even with the recognized, five-point standard error of measurement, Williams scored over 70 on two of these [six] tests."). That approach is no better than accepting the claim if any score intervals include 70.

[147] W. Joel Schneider, *Principles of Assessment of Aptitude and Achievement*, *in* THE OXFORD HANDBOOK OF CHILD PSYCHOLOGICAL ASSESSMENT 286 (Donald H. Saklofske et al. eds. 2013).

[148] *Hall*, 134 S.Ct. at 1997.

[149] Schneider, *supra* note 147, at 290.

[150] *Id*. (describing a procedure for "treat[ing] each IQ test as a subtest of a much larger 'Mega-IQ Test.'"). Another complication, however, is whether later IQ scores should be reduced to account for a population trend toward higher scores over time (the "Flynn effect"). *E.g.*, Ex parte Cathey, 451 S.W.3d 1, 5–6 (Tex. Crim. App. 2014); Nancy Haydt et al., *Advantages of DSM-5 in the Diagnosis of Intellectual Disability: Reduced Reliance on IQ Ceilings in* Atkins *(Death Penalty) Cases*, 82 UMKC L. REV. 359 (2014).

[151] *Hall*, 134 S.Ct. at 1995–96.

[152] Emily Young, *Intelligence Testing: Accurate or Extremely Biased?*, THE NEUROETHICS BLOG (Sept. 24, 2013), http://www.theneuroethicsblog.com/2013/09/intelligence-testing-accurate-or.html. The

aiming to understate his true score.[153] But the only thing that confidence intervals can protect against is random error.[154] If the test is so flawed in other ways that multiple scores are not usable, then, *a fortiori*, single scores cannot be used. If the possibility of repeated mistakes in test administration introduces serious bias, then serious bias is at least as great a problem in interpreting a less precise single test score.

Given the flimsiness of the Court's comments about multiple scores, the best interpretation of them is that they merely mean that the option of retesting in the Florida law is not sufficient to handle the problem of measurement error. Under *Hall*, multiple scores do not obviate the need to consider measurement error through an appropriate interval estimate, but a statistically sound interval estimate based on multiple scores should be admissible to ascertain whether the defendant's true score is 70 or less.

B. Credible Regions (BCR)

A final issue in the construction of confidence intervals or equivalent decision rules is whether the Constitution prevents a state from going beyond the classical test theory discussed in the *Hall* opinions and briefs.[155] The SEM-IS, SEM-AM, and SEE-IS approaches to measurement error all embody the "frequentist"[156] or "classical"[157] theory of statistical inference. Within this framework, a test-taker's true score is a parameter in a statistical model of random error. The "[p]arameters are fixed, unknown constants," which means that "no useful probability statements can be made about parameters."[158] The best we can do is postulate values for the true score and see how often decisions about the true scores of many defendants, based strictly on the observed score (or multiple scores), would be right or wrong. Suppose, for example, that every capital offender has a true score of 70 or less. If the observed scores are normally distributed with the estimated standard error, then the 95% SEM-IS rule would keep the expected rate of false alarms (decisions wrongly preventing a defendant from demonstrating disability with evidence of deficits in adaptive functioning) to no more than 2.5%. And, in no case will we misclassify a nondisabled offender as

---

DSM-5 specifies that "[i]nstruments must be normed for the individual's sociocultural background and native language." DSM-5, *supra* note 20 (Diagnostic Features). *But see* Sanger, *supra* note 10 (criticizing post-*Hall* efforts by some experts to infer higher true IQ scores).

[153] *Hall*, 134 S.Ct. at 2011 n. 13 (Alito, J., dissenting).

[154] KAYE ET AL., *supra* note 107, § 12.6.4; Kaye & Freedman, *supra* note 106.

[155] An alternative to classical test theory is item response theory, which posits a functional relationship between a test-taker's latent ability and the probability of answering a question correctly. FRANK B. BAKER, THE BASICS OF ITEM RESPONSE THEORY (2d ed. 2001). This article deals solely with the classical test theory that the *Hall* Court invoked.

[156] LARRY WASSERMAN, ALL OF STATISTICS 175 (2004).

[157] VIC BARNETT, COMPARATIVE STATISTICAL INFERENCE 123 (3d ed. 1999).

[158] WASSERMAN, *supra* note 156, at 175.

potentially disabled — because there are no such offenders. But suppose that 50% of capitally-convicted offenders have true scores of exactly 70 and that 50% have true scores of 71. No more than approximately 2.5% of the half with IQs of 70 will be misjudged, but about 41% of the half with IQs of 71 will be deemed disabled (if there is sufficient evidence of deficient adaptive functioning) even though not one of them truly qualifies for the exemption.[159]

Calculations of this sort allow us to gauge the performance of any decision rule — its operating characteristics — conditional on the assumed values for the true scores of the offenders. But they cannot tell us the probability that is of interest to the law — the chance that an offender has a true score of 70 or less — or even the proportion of those who are expected to have such low true scores given a defendant's observed score. With information on the distribution of true IQ scores among convicted capital offenders, however, it is feasible to estimate the latter quantities. Suppose a jurisdiction has recorded the IQ scores of convicted capital defendants over the past decade and that these scores are approximately normally distributed with mean 70 and standard deviation 8. This distribution also describes the distribution of true scores. Some defendants' true scores will be above their measured ones, but others will be below. Such differences will wash out, and the overall picture of true scores will mirror that of the measured ones. It will be centered at 70 and spread out so as to have most scores between 65 and 75.[160]

If this historical pattern applies to the next defendant, we should be suspicious of a measured score x like 80 (or 60) that is in a tail of the prior distribution of true scores τ. If we knew nothing of the past, our best guess for the defendant's true score would be the measured one,[161] but we ought to factor in the reality that low and high true scores are the exception and hence are less likely to be the explanation for the extreme score. A better estimate for this defendant's true score τ would be closer to the middle of the pack of true scores — not at the very middle, but somewhere between the observed score x and the average of the previously encountered scores. Moreover, because we are working with more information than just the one newly observed score x, we can have more confidence in the blended estimate than in either the prior mean or the new observation as an estimate of τ. Thus, the interval estimate for the true score can be narrower than the classical CI.

The mathematical recipe for blending the prior distribution of true scores with a newly observed score is known as Bayes' rule.[162] Applying it to a normal prior distribution and an observed

---

[159] The area under the normal curve with mean 71 and standard deviation 2.16 in the region below 70.5 is 0.41.

[160] Almost 51% of the area under the normal curve with mean 70 and standard deviation 8 falls into the region from 64.5 to 75.49.

[161] *See supra* note 86. An frequentist adjustment is appropriate when using the SEE rather than the SEM. *See supra* Part III.F.

[162] *See, e.g.,* BARNETT, *supra* note 157, at 201–50; JEFF GILL, BAYESIAN METHODS: A SOCIAL AND BEHAVIORAL SCIENCES APPROACH 15–18 (3d ed. 2015). Discussions of Bayes' rule in the legal literature began with Michael 0. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970). For a discussion of its judicial acceptance, see KAYE ET AL., *supra* note 107, § 12.8.5.

score with measurement error yields a "Bayesian credible region"[163] (BCR) that has the properties noted above — integrating the prior information shifts the estimate toward the historical mean and produces a narrower interval.[164] BCRs could be used in the same way as CIs to determine whether the credible range of true scores covers a true-score cutoff such as 70. Or, more directly, one can look at the posterior true-score distribution and easily compute whether the portion below this cutoff is sufficiently small to reject the claim of a disability.[165] By way of illustration, if the empirical prior distribution were normal with mean 70 and standard deviation 8, then individuals with a measured score of $x \geq 73$ would have no more than a 10% risk of having a true score $\tau \leq 70$.[166] If 10% is an "unacceptable risk," a higher cut-score for measured IQs would have to be selected. A state willing to tolerate only a 1% risk would use a cut-score of at least 76 in this particular illustration.[167] With modern computational methods, Bayes' rule can be applied to almost any prior distribution, not just the normal ones mentioned here. This form of inference is well established in statistics,[168] psychometrics,[169] and social science.[170] The Supreme Court once characterized it as a "more precise method" than classical hypothesis testing.[171] The approach would enable a state to attain a probability of, say, 95% for a correct decision with respect to the true-score cutoff of 70, as the majority in *Hall* seemed to desire or demand.[172]

Summary and Conclusion

---

[163] PAUL H. GARTWAITE ET AL., STATISTICAL INFERENCE 154 (2d ed. 2002). It also is called a "highest posterior density interval." G. A. YOUNG & R. L. SMITH, ESSENTIALS OF STATISTICAL INFERENCE 30 (2005).

[164] *See, e.g.,* ANDREW GELMAN ET AL., BAYESIAN DATA ANALYSIS (2d ed. 2004). For examples with IQ scores, see Kaye, *supra* note 10.

[165] Kaye, *supra* note 10.

[166] *Id*.

[167] *Id*.

[168] For a sampling of the latest generation of textbooks devoted entirely to the subject, see GELMAN ET AL., *supra* note 164; GILL, *supra* note 162; SIMON JACKMAN, BAYESIAN ANALYSIS FOR THE SOCIAL SCIENCES (2009); JOHN K. KRUSCHKE, DOING BAYESIAN DATA ANALYSIS: A TUTORIAL WITH R, JAGS, AND STAN 22 (2d ed. 2015). A popular history is SHARON BERTSCH MCGRAYNE, THE THEORY THAT WOULD NOT DIE (2011).

[169] *E.g.*, Melvin R. Novick, *Bayesian Methods in Psychological Testing*, 1969 ETS RESEARCH BULLETIN SERIES 1; Hariharan Swaminathan & Janice A. Gifford, *Bayesian Estimation in the Rasch Model*, 7 J. EDUC. STAT. 175 (1982).

[170] *E.g.,* GILL, *supra* note 162; JACKMAN, *supra* note 10; Simon Jackman, *Bayesian Analysis for Political Research*, 7 ANN. REV. POL. SCI. 483 (2004).

[171] Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 312 n.17 (1977).

[172] *Id*.

The *Hall* Court holds that (1) a state can categorically deny the intellectual-disability exemption from capital punishment to individuals whose measured IQ scores are above some cut-score, and (2) this lowest permissible cut-score must be greater than 70 (on a test normed to have a mean of 100 and a standard deviation of 15). Justice Kennedy's majority opinion further suggests that in choosing a cut-score, a state must attend in some manner to the standard error of measurement (SEM), but the opinion does little more than gesture to publications on psychometrics, IQ scores, and intellectual disability that emphasize the importance of recognizing imprecision in IQ scores. To provide more specific guidance, this article has presented four procedures for coping with the one aspect of "measurement error"[173] — namely, random errors in measuring IQ scores — that the Court invoked to invalidate Florida's "rigid rule."[174] The four procedures flow from the fundamental distinction between a test-taker's unknown true score and a measured score. They all seek to control the "unacceptable risk"[175] that an individual with a measured score has a true score that is less than or equal to the maximum true score consistent with a diagnosis of disability. The first three procedures, derived from classical test theory, accomplish this goal indirectly, if at all. The fourth attends directly to the most pertinent probability.

The first procedure (SEM-IS), is the one described, with different degrees of precision and accuracy, in the two opinions. It uses an overall SEM derived from the estimated "reliability" for a specific test and population of test-takers to construct an approximate 95% (or other) confidence interval. If the CI exceeds the targeted true score, the state can deny the *Atkins* exemption without regard to deficits in adaptive functioning. As the dissent notes, the procedure can be applied with any desired confidence coefficient, although the desired level of "confidence" does not have the meaning the dissent (and probably the majority) ascribed to it.

Second, I described an equivalent hypothesis-testing procedure that defines the state's cut-score for the measured IQ as a legislatively targeted true IQ score (such as 70) plus some number of SEMs. Using this SEM-adjusted score (SEM-AS) to reject claims of disability, this procedure tests whether an observed score is statistically significantly larger than 70. It keeps the risk of categorically rejecting a disability claim from a defendant with a true IQ score less than or equal to 70 to a predetermined maximum. This is exactly what the SEM-IS rule does. But without data on the distribution of true scores among defendants, it is impossible to say how well these procedures limit the risk that individuals with measured scores greater than 70 have true scores less than or equal to 70.[176]

The third procedure uses an estimator for true IQ scores based on the distribution of those scores for all individuals with a defendant's test score. Instead of the Court's SEM, it uses a "standard error of estimate" (SEE) that is larger for scores that are far from the population mean. The

---

[173] *Hall*, 134 S.Ct. at 1994.

[174] *Id*. at 1990, 2001.

[175] *Id*. at 1990.

[176] The professional guidelines to which Justice Kennedy deferred use the 95% CI, which is equivalent to the 0.05 two-sided significance level, although the choice is largely conventional.

SEE-IS procedure also shrinks the estimated true score toward the mean.[177]

The fourth and final approach to handling measurement error goes beyond the classical test theory on which the first three were based. It would permit a state to use recent data on the IQ scores of convicted capital offenders together with test-reliability data to estimate the probability that a defendant with a measured IQ has a true score in any desired interval. The Bayesian credible region (BCR) for the highest posterior density could be inspected to see if it contains any true IQ scores that would permit defendant's claim of disability to proceed. The meaning of the region would be clear because, unlike the frequentist CI, the BCR does describe probabilities for true scores. Alternatively, the probability of a true score in the qualifying range could be computed directly from the posterior distribution of true scores. This quantity, which is what Justice Alito thought he was computing. Although it is the quantity of the most legal interest, it cannot be derived from the CI.

In short, *Hall* requires legislatures and courts to take heed of the statistical principles and methods that should guide the interpretation of IQ scores by psychologists and psychiatrists. The essence of the case is the Court's refusal to countenance "an unacceptable risk that persons with intellectual disability will be executed."[178] Understandably, the opinions do not strive to be comprehensive in their discussions of how to achieve this result. Unfortunately, to the extent that the Justices do articulate details of statistical procedures for quantifying the uncertainty of psychological measurements, the dicta are not always dependable.[179] By offering some corrections and elaborations, and by demonstrating that there is more than one statistically acceptable way to try to control the risk of error, this article lays the groundwork for more informed use of the statistical properties of IQ scores in adjudicating claims of intellectual disability.

---

[177] The Court apparently was not aware of the virtues of the SEE, perhaps because its reading of the psychometrics literature was limited to rather basic texts.

[178] The opinion opens with two sentences describing the Florida law and its background and the announcement that "[t]his rigid rule, the Court now holds, creates an unacceptable risk that persons with intellectual disability will be executed, and thus is unconstitutional." 134 S.Ct. at 1990.

[179] *Hall* is by no means the only case in which the Justices' efforts to describe statistical inference or to rely on statistical reasoning have not brought them glory. *See* KAYE ET AL., *supra* note 107, § 7.3.2(c)(1), at 323–24 (criticizing opinions in Barefoot v. Estelle, 463 U.S. 880 (1983)); David H. Kaye, *Trapped in the* Matrixx: *The U.S. Supreme Court and the Need for Statistical Significance*, 39 PROD. SAFETY & LIAB. REP. (BNA) 1007 (2011) (identifying potentially problematic dicta in *Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309 (2011)); David H. Kaye, *And Then There Were Twelve: The Supreme Court, Statistical Reasoning, and the Size of the Jury*, 68 CAL. L. REV. 401 (1980) (criticizing Justice Blackmun's opinion in *Ballew v. Georgia*, 463 U.S. 880 (1983)); Richard O. Lempert, *Uncovering "Nondiscernible" Differences: Empirical Research and the Jury-Size Cases*, 73 MICH. L. REV. 643 (1975) (criticizing the majority opinion in *Williams v. Florida*, 399 U.S. 78 (1970)).