UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF NEW YORK
----------X

UNITED STATES OF AMERICA

-against-

RONELL WILSON,

Defendant.

MEMORANDUM & ORDER

04-CR-1016 (NGG)

----------X

NICHOLAS G. GARAUFIS, United States District Judge.

The execution of those who are mentally retarded violates both the Federal Death Penalty Act and the Eighth Amendment. See 18 U.S.C. § 3596(c); Atkins v. Virginia, 536 U.S. 304, 321 (2002). Defendant Earl Ronell Wilson, a convicted murderer of two undercover police officers, claims that he is mentally retarded and therefore ineligible to receive the death penalty. For the reasons that follow, he is incorrect.

I. BACKGROUND[1]

In 2003, Wilson murdered two undercover NYPD detectives who were posing as gun purchasers. (See Second Superseding Indictment (Dkt. 179) ¶¶ 7, 9.) He was tried in this court for capital-eligible crimes. (Trial Tr. (Dkts. 362-404).) The jury convicted Wilson and voted to impose the death penalty. (Jury Verdict (Dkt. 351).) The court accordingly sentenced Wilson to death. (Judgment (Dkt. 407).) Wilson appealed, and the Second Circuit affirmed Wilson's convictions but vacated his death sentence on constitutional grounds and remanded to this court for retrial of his penalty phase. United States v. Whitten, 610 F.3d 168 (2d Cir. 2010).

[1] The court will discuss only the background pertinent to the issues it addresses in this opinion. Additional background can be found in the Second Circuit's decision in this case. See United States v. Whitten, 610 F.3d 168, 173-77 (2d Cir. 2010).

After the Second Circuit's mandate issued, Wilson requested "a pretrial hearing to determine whether he is a person with mental retardation" and thus ineligible for the death penalty under the Eighth Amendment and the Federal Death Penalty Act (the "Atkins claim"). (Dkt. 614.) The court granted this request and set a schedule for exchange of expert information, motions related to the Atkins claim, and an evidentiary hearing (the "Atkins hearing"). (See Feb. 2, 2012, Order (Dkt. 618).)

Wilson provided notice of his intent to call four mental health experts at the Atkins hearing: (1) John Olley, Ph.D., a psychologist; (2) Bruce Shapiro, M.D., a developmental pediatrician; (3) Joette James, Ph.D., a neuropsychologist; and (4) George Woods, Jr., a licensed physician. (Mar. 7, 2012, Def. Ltr. (Dkt. 637).) The Government stated that it would call three experts at the hearing: (1) Robert Denney, Psy.D., a neuropsychologist; (2) Robert Mapou, Ph.D., a neuropsychologist; and (3) Raymond Patterson, M.D., a psychiatrist. (Mar. 7, 2012, Gov't Ltr. (Dkt. 638); Apr. 6, 2012, Gov't Ltr. (Dkt. 676); May 2, 2012, Gov't Ltr. (Dkt. 697).) Since then, the parties and their experts have conducted extensive discovery and testing in preparation for the Atkins hearing. The court has also issued two opinions in response to motions filed by the parties regarding the scope of discovery. See United States v. Wilson, No. 04-CR-1016 (NGG), 2012 WL 3890951, at *4-8 (E.D.N.Y. Sept. 7, 2012); United States v. Wilson, No. 04-CR-1016 (NGG), 2012 WL 6962982, at *6-16 (E.D.N.Y. June 22, 2012).

On September 7, 2012, the parties exchanged expert reports. (See Sept. 7, 2012, Def. Ltr. (Dkt. 868).) Each of the Government's experts opined that Wilson is not mentally retarded. (Denney Rep. (Dkt. 956) at 48; Patterson Rep. (Dkt. 957) at 18; Mapou Rep. (Dkt. 958) at 35.) Each of Wilson's experts opined that he is mentally retarded. (James Rep. (Dkt. 959) at 1, 17; Olley Rep. (Dkt. 960) at 28; Shapiro Rep. (Dkt. 961) at 2, 22; Woods Rep. (Dkt. 962) at 29.)

The court held the Atkins hearing over nine days in November and December 2012. (See Minute Entries (Dkts. 950-55, 976-78).) It heard testimony from all seven of the experts mentioned above and four other witnesses. (See Atkins Hr'g Tr. ("Tr.").) The parties submitted briefing on the Atkins claim after the hearing. (Def. Mem. (Dkt. 982); Gov't Mem. (Dkt. 983);[2] Def. Reply (Dkt. 999).)

II. STANDARD FOR MENTAL RETARDATION

Two provisions of law forbid federal courts from imposing a death sentence upon a person who is mentally retarded. First, the Federal Death Penalty Act ("FDPA"), originally enacted by Congress in 1988 and amended in 1994, provides that a "sentence of death shall not be carried out upon a person who is mentally retarded." 18 U.S.C. § 3596(c). Second, the execution of mentally retarded individuals violates the Eighth Amendment's ban on "cruel and unusual punishments." Atkins v. Virginia, 536 U.S. 304, 321 (2002); see U.S. Const. amend. VIII ("Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted.").

In Atkins, the Supreme Court found that, due to the relatively recent legislative efforts of several states, "a national consensus ha[d] developed against" the execution of mentally retarded offenders. Atkins, 546 U.S. at 316. Such executions were therefore inconsistent with "'the evolving standards of decency that mark the progress of a maturing society'"—the guiding principle of the Eighth Amendment. Id. at 311-12 (quoting Trop v. Dulles, 356 U.S. 86, 101 (1958)). The Court concluded that although the intellectual deficiencies of mentally retarded criminals did "not warrant an exemption from *criminal sanctions*"—including life imprisonment—such criminals "should be categorically excluded from *execution*" for two main

[2] Two days after filing its brief, the Government filed an "amended" brief that corrects a few formatting issues with the original brief. (See Dkt. 984.) All citations to the Government's brief refer to the first version. (See Dkt. 983.)

reasons. Id. at 318 (emphases added). First, there was a "serious question" as to whether the execution of mentally retarded offenders would serve the deterrence or retribution justifications of the death penalty. Id. at 318-319. Second, there was an enhanced risk in the case of mentally retarded offenders "that the death penalty w[ould] be imposed in spite of factors which may call for a less severe penalty," both because of "the possibility of false confessions" by mentally retarded defendants and because of the "lesser ability of mentally retarded defendants to make a persuasive showing of mitigation." Id. at 320.

It is therefore clear that this court may not sentence a mentally retarded criminal to death, but that is where most of the clarity ends. The difficult task is deciding which persons qualify as "mentally retarded" under the FDPA and Atkins—an issue of first impression in this Circuit.

A. Sources of the Definition

Neither the FDPA nor Atkins mandates a particular definition of mentally retardation. The FDPA provides simply that "mentally retarded" persons may not be executed. 18 U.S.C. § 3596(c); see also Garcia Briseno v. Dretke, No. 05-CV-08, 2007 WL 998743, at *10 n.8 (S.D. Tex. Mar. 29, 2007). And Atkins expressly left "to the States the task of developing appropriate ways to enforce the constitutional restriction upon their execution of sentences." 536 U.S. at 317 (alterations omitted); see also Bobby v. Bies, 129 S. Ct. 2145, 2150 (2009) ("Our opinion [in Atkins] did not provide definitive procedural or substantive guides for determining when a person who claims mental retardation 'will be so impaired as to fall [within Atkins' compass].'" (second alteration in original) (quoting Atkins, 536 U.S. at 317)). Atkins noted, however, that although state "statutory definitions of mental retardation [we]re not identical, [they] generally conform[ed] to [ ] clinical definitions" promulgated by two groups: (1) the American Association on Mental Retardation ("AAMR"), which has since changed its name to the

American Association on Intellectual and Developmental Disabilities ("AAIDD"); and (2) the American Psychiatric Association ("APA"). Id. at 308 n.3, 317 n.22. The Court further noted that these clinical definitions "require not only subaverage intellectual functioning, but also significant limitations in adaptive skills such as communication, self-care, and self-direction that became manifest before age 18." Id. at 318. In short, Atkins declined to mandate a definition of mental retardation but left it to the states to define the term, while noting that existing state definitions generally conformed to the clinical definitions set forth by the AAMR and the APA.

Atkins did not hold that *federal* courts are bound to apply the mental retardation definitions of the particular states in which they are located, nor does the FDPA contain any such mandate. Federal courts that have decided cases involving both Atkins and FDPA claims have taken inconsistent approaches in this regard: some have applied their forum state's law, see, e.g., United States v. Cisneros, 385 F. Supp. 2d 567, 571-72 (E.D. Va. 2005) (applying Virginia law), while others have made no mention of their state's law and applied only clinical definitions of mental retardation, see, e.g., United States v. Hardy, 762 F. Supp. 2d 849 (E.D. La. 2010) (no mention of Louisiana law); United States v. Davis, 611 F. Supp. 2d 472 (D. Md. 2009) (no mention of Maryland law).[3] This court will consider New York law in determining the definition of mental retardation, while noting that: (1) Atkins does not explicitly require that the court be *bound* by New York law; and (2) even if it did, an application of New York law would, as discussed below, ultimately lead the court to rely primarily upon clinical definitions of the term.

[3] In cases involving petitions for writ of habeas corpus pursuant to 28 U.S.C. § 2254, federal courts have of course applied state law, as these claims required the court to review a state court's determination of the petitioner's Atkins claim (and did not involve the FDPA). See, e.g., Williams v. Mitchell, No. 09-CV-2246, 2012 WL 4505774, at *35 (N.D. Ohio Sept. 28, 2012) (applying Ohio law in a § 2254 case); Thomas v. Allen, 614 F. Supp. 2d 1257, 1262-63 (N.D. Ala. 2009) (same with Alabama law).

New York has been without the death penalty since 2004, when the New York Court of Appeals held that the State's capital sentencing statute violated its Constitution. See People v. LaValle, 3 N.Y.3d 88 (2004). This statute is, however, still on the books for the most part. See New York Criminal Procedure Law ("C.P.L.") § 400.27. As before LaValle, the statute requires a court to side aside a defendant's capital sentence if it finds that the defendant is mentally retarded, with certain exceptions. See id. § 400.27(12). The statute provides further that "'mental retardation' means significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior which were manifested before the age of eighteen." Id. § 400.27(12)(e). Atkins cited this statute in support of its finding that a national consensus had developed against the execution of mentally retarded individuals. See 536 U.S. at 314 & n.13; see generally People v. Smith, 751 N.Y.S.2d 356, 357 (N.Y. Sup. Ct. 2002).

New York's statute, however, provides little guidance as to the definition of mental retardation for three reasons. First, although the statute is still technically in force, it has been effectively rendered a nullity by the invalidation of New York's death penalty scheme, and thus can be considered at most only a weak expression by the State of the definition of mental retardation for Atkins purposes. Second, even if the statute could inform the definition of mental retardation under Atkins, it likely would not affect the definition under the FDPA, which independently forbids the execution of mentally retarded offenders. See 18 U.S.C. § 3596(c). Third, the definition in New York's statute is essentially identical to the clinical definitions discussed below, and neither the statute itself nor any New York case law provides guidance beyond the statute's definitional statement. Indeed, the language of the statute tracks very closely with a 1983 definition propounded by the AAIDD (then known as the American Association on Mental Deficiency). See American Association on Mental Deficiency,

Classification in Mental Retardation 1 (8th ed. 1983) ("Mental retardation refers to significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior, and manifested during the developmental period.").

For these reasons, the court relies largely on the clinical definitions of mental retardation promulgated by the AAIDD and the APA, the two leading authorities on the subject. These authorities were cited favorably in Atkins, and nothing in either the FDPA or New York law prevents the court from relying upon them. Most federal courts have taken the same approach when deciding Atkins cases. See, e.g., United States v. Smith, 790 F. Supp. 2d 482, 485-86 (E.D. La. 2011); United States v. Lewis, No. 08-CR-404 (SO), 2010 WL 5418901, at *5, *23 (N.D. Ohio Dec. 23, 2010); Hardy, 762 F. Supp. 2d at 854; Davis, 611 F. Supp. 2d at 474.

The court emphasizes, however, that "psychology informs, but does not determinatively decide, whether an inmate is exempt from execution." Ortiz v. United States, 664 F.3d 1151, 1168 (8th Cir. 2011). Atkins "did not delegate to psychologists the determination of whether an inmate should not face execution." United States v. Bourgeois, No. 02-CR-216, 2011 WL 1930684, at *24 (S.D. Tex. May 19, 2011); see also Hooks v. Workman, 689 F.3d 1148, 1172 (10th Cir. 2012) ("Atkins could have adopted the clinical standard, but explicitly declined to do so."); Clark v. Quarterman, 457 F.3d 441, 445 (5th Cir. 2006) (Atkins "did not dictate that the approach" to defining mental retardation "must track the approach of the [AAIDD] or the APA exactly"); United States v. Candelario-Santana, No. 09-CV-427 (JAF), 2013 WL 101615, at *2 (D.P.R. Jan. 8, 2013) ("Though the clinical standards have informed our analysis, we emphasize that a clinical standard is not a constitutional command." (internal quotation marks omitted)). Instead, while noting the leading clinical definitions of mental retardation, Atkins expressly permitted state legislatures and courts to exercise their own judgments as to the definition of

mental retardation, even if those judgments diverged from those of leading psychologists. See Atkins, 536 U.S. at 317 ("[W]e leave to the States the task of developing appropriate ways to enforce the constitutional restriction upon their execution of sentences." (alterations omitted)); see also id. at 317 n.22 ("The statutory definitions of mental retardation . . . *generally* conform to the clinical definitions . . . ." (emphasis added)). This must logically be true as well in situations like this one in which a federal court must define the term in the absence of significant state legislative or state judicial guidance. See Bourgeois, 2011 WL 1930684, at *24 (Atkins "left the contours of the constitutional protection to the courts"). The court will thus rely heavily upon clinical definitions and expert testimony to determine the definition of mental retardation for capital punishment purposes, but, particularly where these definitions and testimony are ambiguous or conflicting (as they often are in this case), it will apply its own judgment as to the "appropriate ways" to enforce the ultimately *legal* prohibition on executing mentally retarded offenders. Atkins, 536 U.S. at 317.

The court must also decide whether it should rely upon *current* clinical definitions of mental retardation or those that were in place at the time of Atkins. Although Atkins cited the APA definition that is used today, see 536 U.S. at 308 n.3 (citing APA, Diagnostic and Statistical Manual of Mental Disorders 41 (4th ed. 2000) ("DSM-IV-TR")), it cited the 1992 version of the AAIDD's definition, see id. (citing AAMR, Mental Retardation: Definition, Classification, and Systems of Supports 5 (9th ed. 1992)), which the AAIDD has since supplanted with two recent publications, see AAIDD, Intellectual Disability: Definition, Classification, and Systems of Supports (11th ed. 2010) ("AAIDD 2010 Manual"); AAIDD, User's Guide: Intellectual Disability: Definition, Classification, and Systems of Supports (11th ed. 2012) ("AAIDD 2012 User's Guide"). The Government argues that because "these later AAIDD materials were not

contemplated by the Atkins Court" and "fall outside the scope of the 'national consensus' upon which the Supreme Court relied in Atkins," the court should not adopt them as part of the AAIDD's definition of mental retardation. (Gov't Mem. at 18-19; see also id. at 24-26.)

The court disagrees. Contrary to the Government's argument, Atkins did not conclude that there was a national consensus as to the *definition* of mental retardation; it suggested just the opposite. See 536 U.S. at 317 ("To the extent there is serious disagreement about the execution of mentally retarded offenders, it is in determining which offenders are in fact retarded."). What the Court concluded was that there was a national consensus against execution of those offenders that fit within a *given state's* definition of mental retardation, while permitting the states to continue to define the contours of the definition in their own—and differing—ways. See id. The Government cannot seriously dispute that a state would be permitted to define mental retardation according to current clinical definitions as opposed to those existing at the time of Atkins. It logically follows that, in the absence of binding law to the contrary, this court is also permitted to exercise its judgment as to the best interpretation of "mental retardation," even if that interpretation diverges from the understanding of the term at the time Atkins was decided. In any event, surely nothing in the FDPA prevents the court from doing so.

It is also important to note that the Government's approach would be very difficult (if not impossible) to apply in practice. For example, if Atkins requires the court to apply only the clinical standards in place at the time it was decided, does that mean the court prohibited from considering intelligence tests developed after Atkins? The Government apparently does not think so, because its own expert, Dr. Denney, administered an intelligence test on Wilson that was published in 2008, six years after Atkins. (See Denney Rep. at 40.) Moreover, as will be discussed in Part III, clinical judgment is essential to the interpretation of intelligence testing.

The Government's approach may require clinicians to set aside much of their training in post-Atkins psychological standards and to train themselves (for Atkins purposes alone) in the outdated standards existing in 2002. Atkins should not be read to require this result.[4]

Thus, because the AAIDD 2010 Manual reflects the AAIDD's view of the current best practices in the field, the court will rely upon this edition. Such reliance is permissible under Atkins and the FDPA, sensible as a practical matter, and consistent with the approach of other federal courts. See, e.g., United States v. Northington, No. 07-CR-550-05, 2012 WL 4024944, at *3 (E.D. Pa. Sept. 12, 2012) ("Northington II"); Smith, 790 F. Supp. 2d at 484; Bourgeois, 2011 WL 1930684, at *23 n.27; Lewis, 2010 WL 5418901, at *8; Hardy, 762 F. Supp. 2d at 854 n.5.[5]

B. Clinical Definitions of Mental Retardation

The definitions of mental retardation set forth by the AAIDD and the APA are "essentially identical." Davis, 611 F. Supp. 2d at 475; see also Ortiz, 664 F.3d at 1158; United States v. Northington, No. 07-CR-550-05, 2012 WL 2873360, at *2 n.6 (E.D. Pa. July 12, 2012) ("Northington I"); Lewis, 2010 WL 5418901, at *5; see generally United States v. Nelson, 418 F. Supp. 2d 891, 894-95 (E.D. La. 2006) (explaining the minor differences between the definitions and noting that they "do not appear to conflict").

According to the APA, a diagnosis of mental retardation requires:

---

[4] The Government quotes the Supreme Court's statement that "'[n]ot all people who claim to be mentally retarded will be so impaired as to fall within the range of mentally retarded offenders about whom there is a national consensus.'" (Gov't Mem. at 24 (quoting Atkins, 536 U.S. at 317).) The Government apparently interprets this statement to mean that every aspect of the mental retardation definition must be consistent with a national consensus in order to exempt a defendant from execution. Once again, this interpretation is inconsistent with the Supreme Court's express allowance of different definitions of mental retardation in different states. See Atkins, 536 U.S. at 317. The portion of Atkins quoted by the Government, although somewhat ambiguous, appears to stand for the unsurprising proposition that a person cannot simply "*claim* to be mentally retarded" to obtain an exemption from the death penalty. Id. (emphasis added).

[5] The court is aware of no case in which a court has considered itself bound to apply outdated clinical standards in making an Atkins determination.

> A. Significantly subaverage intellectual functioning: an IQ of approximately 70 or below on an individually administered IQ test . . . .
> B. Concurrent deficits or impairments in present adaptive functioning (i.e., a person's effectiveness in meeting the standards expected for his or her age by his or her cultural group) in at least two of the following areas: communication, self-care, home living, social/interpersonal skills, use of community resources, self-direction, functional academic skills, work, leisure, health and safety.
> C. The onset is before 18 years of age.

DSM-IV-TR at 49.

The AAIDD defines mental retardation (which it now calls "intellectual disability" or "ID"[6]) as follows: "Intellectual disability is characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills. This disability originates before 18." AAIDD 2010 Manual at 1.

Putting these two clinical standards together, the definition of mental retardation has three "uniformly accepted" requirements, Bourgeois, 2011 WL 1930684, at *24, which the court will at times refer to as "prongs one, two, and three." These requirements are: (1) significantly subaverage intellectual functioning; (2) significant deficits in adaptive behavioral skills; and (3) onset of the condition before age eighteen. See AAIDD 2010 Manual at 7, 27, 41; DSM-IV-TR at 49; Taylor v. Quarterman, 498 F.3d 306, 307 (5th Cir. 2007); Northington II, 2012 WL 4024944, at *3; Davis, 611 F. Supp. 2d at 475; cf. Atkins, 536 U.S. at 318. The three prongs are each "indispensable" to a finding of mental retardation. Blue v. Thaler, No. 05-CV-2726 (H), 2010 WL 8742423, at *9 (S.D. Tex. Apr. 19, 2010), aff'd, 665 F.3d 647 (5th Cir. 2011); see also AAIDD 2010 Manual at 7 (the three elements are each "essential").

---

[6] "In recent years, the field of psychology has favored the use of the clinical designation 'intellectual disability' instead of 'mental retardation.'" Northington I, 2012 WL 2873360, at *2 n.6. These terms are synonymous. See AAIDD 2010 Manual at 12 ("[T]he term ID covers the same population of individuals who were diagnosed previously with mental retardation . . . . [E]very individual who is or was eligible for a diagnosis of mental retardation is eligible for a diagnosis of ID."); see also Ricci v. Okin, 770 F. Supp. 2d 438, 440 n.2 (D. Mass. 2011) ("The phrase 'intellectual disability' or variations thereof is synonymous with 'mental retardation.'"). Since the FDPA and Atkins use the term "mentally retarded," the court will do so as well. It will, however, refer to the aforementioned organization by its current acronym, the AAIDD.

The third prong—onset before the age of eighteen—bears clarification because it is essentially a prerequisite to satisfying the first two prongs. To be mentally retarded, both an individual's significantly subaverage intellectual functioning and his significant deficits in adaptive behavioral skills must become manifest before the age of 18.[7] See Blue, 2010 WL 8742423, at *9 (definition requires "(1) substantial limitations in intellectual functioning; (2) significant limitations in adaptive area skills; and (3) manifestation of *those limitations* before age 18" (emphasis added)); see also Atkins, 536 U.S. at 318 ("[C]linical definitions require not only subaverage intellectual functioning, but also significant limitations in adaptive skills such as communication, self-care, and self-direction *that* became manifest before age 18." (emphasis added)). Moreover, in deciding an Atkins claim, the court must determine whether the defendant "was mentally retarded *at the time of the crime*." Hardy, 762 F. Supp. 2d at 881 (emphasis added); see also Holladay v. Allen, 555 F.3d 1346, 1353 (11th Cir. 2009) ("Though the factors state that the problems had to have manifested themselves before the defendant reached the age of eighteen, it is implicit that the problems also existed at the time of the crime." (internal quotation marks omitted)). Thus, mental retardation must "be diagnosed, if it is to be diagnosed at all, retrospectively in every sense of the word." Hardy, 762 F. Supp. 2d at 881.

C. Additional Legal Principles

As noted above, clinical definitions of mental retardation do not provide the full picture for an Atkins case. Two important general legal principles govern the court's analysis.

First, whether an individual is mentally retarded "is a question of fact, and not a mixed question of law and fact." Clark, 457 F.3d at 444; see also Ortiz, 664 F.3d at 1164; Walker v.

[7] This does not mean that a defendant must be *diagnosed* with mental retardation before the age of eighteen, only that the disability's defining symptoms must have manifested themselves before the age of eighteen. (See Shapiro Rep. at 12.) See also AAIDD 2010 Manual at 27 ("[D]isability does not necessarily have to have been formally identified, but it must have originated during the developmental period.").

Kelly, 593 F.3d 319, 323 (4th Cir. 2010). The *standard* for whether someone is mentally retarded and ineligible for the death penalty under the Eighth Amendment and FDPA is a legal matter (as discussed above), but "the ultimate issue of whether [Wilson] is, in fact, mentally retarded" is for the court to decide as a factual matter, "based upon all of the evidence and determinations of credibility." In re Briseno, 135 S.W.3d 1, 9 (Tex. Crim. App. 2004).

Second, the court must decide the burden of proof for Atkins/FDPA claims. Although neither Atkins nor the FDPA addresses this issue, the vast majority of courts to address it have held that the defendant bears the burden of proof by a preponderance of the evidence. See, e.g., Northington II, 2012 WL 4024944, at *3; Smith, 790 F. Supp. 2d at 484; Bourgeois, 2011 WL 1930684, at *46; Lewis, 2010 WL 5418901, at *4; Hardy, 762 F. Supp. 2d at 851; Davis, 611 F. Supp. 2d at 474; Thomas v. Allen, 614 F. Supp. 2d 1257, 1296 (N.D. Ala. 2009); Nelson, 419 F. Supp. 2d at 894. But see, e.g., People v. Vasquez, 84 P.3d 1019, 1023 (Colo. 2004) (upholding a Colorado statute requiring proof by the defendant by clear and convincing evidence); Head v. Hill, 277 Ga. 255, 261 (2003) (upholding a Georgia statute requiring proof by the defendant beyond a reasonable doubt).[8] This is also the standard set forth in New York's capital statute. See C.P.L. § 400.27(12)(a) ("[T]he defendant has the burden of proof by a preponderance of the evidence that he or she is mentally retarded."). Both parties agree that this burden applies here (see Def. Mem. at 1; Gov't Mem. at 49), and the court finds no compelling reason to apply a more stringent burden of proof. It therefore adheres to the majority and New York position: Wilson will have the burden of proving that he is more likely than not mentally retarded.

[8] The court is aware of no case in which a court has placed the burden of proving mental retardation on the government. Cf. In re Briseno, 135 S.W.3d at 12 ("The issue of mental retardation is similar to affirmative defenses such as insanity, incompetency to stand trial, or incompetency to be executed, for which the . . . burden of proof [has been placed] upon a defendant . . . .").

* * * * *

In sum, both the FDPA and the Eighth Amendment forbid the court from imposing a death sentence upon a person who is mentally retarded. A person is mentally retarded only if he satisfies three necessary requirements: (1) significantly subaverage intellectual functioning; (2) significant deficits in adaptive behavioral skills; and (3) onset of those limitations before the age of 18. In developing the nuances of these requirements, the court will rely heavily upon modern clinical definitions of mental retardation, particularly those of the APA and the AAIDD. At the same time, the definition of mental retardation is ultimately a legal matter, and so the court may—and will—exercise its own judgment as to the appropriate definition of mental retardation in the Atkins/FDPA context. Once the court expounds upon the legal definition of mental retardation, it must decide as a factual matter whether Wilson himself is mentally retarded, an issue on which Wilson bears the burden of proof by a preponderance of the evidence.

III. INTELLECTUAL FUNCTIONING

With these principles in mind, the court turns to the first prong of the mental retardation definition: significantly subaverage intellectual functioning. The court will begin by discussing intellectual functioning generally, with a focus on some of the especially difficult and disputed issues involved in analyzing it. (See Part III.A.) Resolution of these issues will then guide the court's analysis of Wilson's own intellectual functioning. (See Part III.B.)

A. Intellectual Functioning in General

The AAIDD has stated that "intellectual functioning is currently best conceptualized and captured by a general factor of intelligence," and defines "intelligence" as "a general mental ability" that "includes reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, learning quickly, and learning from experience." AAIDD 2010

Manual at 31. Intellectual functioning is primarily evaluated using standardized tests that measure a person's "Intelligence Quotient," or "IQ." See id. ("Although far from perfect, intellectual functioning is currently best represented by IQ scores when they are obtained from appropriate, standardized and individually administered assessment instruments."); Hardy, 762 F. Supp. 2d at 875 ("Both the APA and AAMR/AAIDD indicate that a diagnosis of mental retardation should be made based on IQ test results where it is possible to perform such a test."); Thomas, 614 F. Supp. 2d at 1264. At the same time, the AAIDD makes clear that IQ scores themselves do not tell the whole story about someone's intelligence; rather, "one needs to use clinical judgment" to interpret those scores and other relevant information.[9] AAIDD 2010 Manual at 35.

The most widely-accepted IQ tests in the United States are the Wechsler Intelligence Scales, which include the Wechsler Intelligence Scale for Children ("WISC") and the Wechsler Adult Intelligence Scale ("WAIS"). Each Wechsler test is composed of several subtests, some of which evaluate a person's "verbal" abilities and some a person's "performance" abilities. A person's IQ is calculated by adding together the number of points earned on all of the subtests and then using a mathematical formula to convert this raw score into an overall score, called the "full scale IQ." In addition to the full scale IQ, the tests also produce a "verbal IQ" and a "performance IQ," which are based solely on the subtests relating to those particular skills.

The Wechsler tests are "standardized" instruments, meaning that during their design phases, they were administered to a large, representative sample of the population in order to predict the distribution of results that the general population would likely obtain. See Thomas, 614 F. Supp. 2d at 1264. The results obtained by the representative sample were then analyzed

[9] The AAIDD defines "clinical judgment" as "a special type of judgment rooted in a high level of clinical expertise and experience and judgment that emerges directly from extensive training, experience with the person, and extensive data." AAIDD 2010 Manual at 29.

for the purposes of creating scaled test scores, a process called "norming." All of the Wechsler tests are normalized so that their average scaled score is 100. The "standard deviation" indicates how far a score is away from that average. It is the baseline against which a person's intellectual deficits (or strengths) are measured, and can be translated into a percentile that indicates a person's relative intelligence within the population. The Wechsler tests are normalized to have a standard deviation of 15 points. These concepts are often depicted with a bell-shaped curve.

Although their definitions are worded somewhat differently, both the AAIDD and the APA define significantly subaverage intellectual functioning by reference to an IQ score approximately two standard deviations below the mean, or 70. The AAIDD defines it as "an 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 2010 Manual at 27. The APA similarly defines it as "an IQ of approximately 70 or below on an individually administered IQ test."[10] DSM-IV-TR at 49.

These definitions are simple enough to state, but they raise several challenging issues.

1. The Standard Error of Measurement and Confidence Intervals

Although both the APA and the AAIDD refer to an IQ score of approximately 70 in their definitions of significantly subaverage intellectual functioning, neither advocates the use of a fixed "cutoff score" for a finding of mental retardation, and the AAIDD explicitly advises against it. See AAIDD 2010 Manual at 40 ("A fixed point cutoff score for ID is not psychometrically justifiable."). Instead, "[b]oth the APA and [AAIDD] direct that [an IQ] test's

[10] The APA categorizes mental retardation as mild, moderate, severe, and profound, with a residual category of "mental retardation, severity unspecified." DSM-IV-TR at 42-44. "Mild" mental retardation, associated with an IQ of 50-55 to 70-75, is the largest segment (about 85%) of those with the disorder; "moderate" mental retardation means an IQ in the range of 35-40 to 50-55; "severe" mental retardation means an IQ of 20-25 to 35-40; and "profound" mental retardation means an IQ below 20-25. Id. at 42-43. The law does not draw a distinction between these categories for the purposes of an Atkins or FDPA claim. See Bourgeois, 2011 WL 1930684, at *25 n.31.

measurement error must be taken into account when interpreting its result." Hardy, 762 F. Supp. 2d at 856; see DSM-IV-TR at 41-42; AAIDD 2010 Manual at 35.

The concept of measurement error is grounded in the idea that each person has a "true" IQ score—the hypothetical score that person would obtain if no error influenced the results of the IQ test. Thomas, 614 F. Supp. 2d at 1269. All IQ tests, however, contain at least some possibility of error, making it impossible to state a test subject's "true" IQ score with certainty. Id. "An IQ score is subject to variability as a function of a number of potential sources of error, including variations in test performance, examiner's behavior, cooperation of the test taker, and other personal and environmental factors." AAIDD 2010 Manual at 36.

The Wechsler tests take measurement errors into account through the use of a mathematical concept known as the "Standard Error of Measurement" ("SEM"). See id.; Thomas, 614 F. Supp. 2d at 1270. The SEM is an index showing the variability of test scores produced by the people forming the normative sample for a particular test. Thomas, 614 F. Supp. 2d at 1270. It is used to calculate a "confidence interval," or a range of scores within which we can be confident to a certain degree that an individual's "true" IQ lies. Id. The confidence interval refers to a percentage corresponding to our degree of confidence that an interval around the obtained IQ score contains the true IQ score. See Wiley v. Epps, 668 F. Supp. 2d 848, 893-94 (N.D. Miss. 2009). For example, the 95% confidence interval for a given IQ score would show the range of scores within which we can be 95% confident that a person's true IQ score falls. This means that "if we could administer the test to that person 100 times (as if new each time), 95 times out of the 100 we would observe a score that is within those confidence bounds." (James Rep. at 4.)

The AAIDD discusses two particular confidence intervals bearing a relatively simple relationship to the SEM: (1) the 66% confidence interval is the range from one SEM below the observed score to one SEM above the observed score; and (2) the 95% confidence interval is the range from two SEMs below the observed score to two SEMs above the observed score. See AAIDD 2010 Manual at 36. For example, the most recent edition of the Wechsler test has an average SEM of 2.3 points. (See Shapiro Rep. at 8.) Thus, if an individual scored a 70 on this test, the 66% confidence interval would be from 67.7 to 72.3 (i.e., 70 $\pm$ 2.3), and the 95% confidence interval would be from 65.6 to 74.6 (i.e., 70 $\pm$ 4.6).

Courts and state legislatures generally have not focused on the precise SEM of particular IQ tests or on the choice of a particular percentage confidence interval. Those courts that have accounted for measurement error generally have found that the SEM for well-standardized IQ tests is approximately 5 points (without distinguishing based on the test edition), and, using a range from one SEM below to one SEM above the observed score (the 66% confidence interval), have set an upper bound of 75 for a finding of mental retardation.[11] See, e.g., Smith, 790 F. Supp. 2d at 490 ("There is [ ] general agreement . . . that a score of 75 should be used as the upper bound of the IQ range describing mild mental retardation."); Bourgeois, 2011 WL 1930684, at *25 ("Because IQ tests typically have a standard error of measurement (also called a 'confidence interval' or 'confidence band'), a base IQ score actually represents a range that could be five points higher or lower. Thus, the psychological profession accepts 75 as a qualifying score for a diagnosis of mental retardation."); Lewis, 2010 WL 5418901, at *8

[11] Some courts have not accounted for the SEM at all, often on the basis of state law. See, e.g., Williams v. Mitchell, No. 09-CV-2246, 2012 WL 4505774, at *35-36 (N.D. Ohio Sept. 28, 2012) (noting that "Ohio law does not mandate consideration of [the] SEM"); Pizzuto v. Blades, No. 05-CV-516 (BLW), 2012 WL 1189908, at *3 (D. Idaho Apr. 6 2012) ("[T]he Idaho Supreme Court seems to have rejected consideration of a standard error of measurement altogether."). The court finds this approach inconsistent with modern best clinical practices, and nothing in New York law precludes it from accounting for the SEM. Cf. C.P.L. § 400.27(12).

("Taking into consideration the SEM of 5 points on either side of 70, an IQ score for intellectual disability falls within the range of 65 to 75."); Hardy, 762 F. Supp. 2d at 857 ("All the experts in this case agree that a score of 75 should be used as the upper bound of the IQ range describing mild mental retardation. Indeed, there is almost universal agreement on this point."); Blue, 2010 WL 8742423, at *9 ("The psychological profession [ ] sets 75 as the base score that may qualify for a diagnosis of mental retardation . . . ."); Wiley, 668 F. Supp. 2d at 893 ("[A]n obtained score of 70 on a Wechsler intelligence test actually represents a range of possible scores between 65 and 75."); Davis, 611 F. Supp. 2d at 475 ("[T]he SEM in IQ assessments is approximately 5 points, therefore raising the operational definition of mental retardation to 75."); see also Atkins, 536 U.S. at 309 n.5 ("It is estimated that between 1 and 3 percent of the population has an IQ between 70 and 75 or lower, which is typically considered the cutoff score for the intellectual function prong of the mental retardation definition."). In addition to any applicable state statutes, these courts have relied upon parts of the DSM-IV-TR and previous publications by the AAIDD that appear to advise the use of a 5-point SEM and an approximate upper bound of 75. See DSM-IV-TR at 41-42 ("It should be noted that there is a measurement error of approximately 5 points in assessing IQ . . . . Thus, it is possible to diagnose Mental Retardation in individuals with IQs between 70 and 75 who exhibit significant deficits in adaptive behavior."); AAMR, Mental Retardation: Definition, Classification, and Systems of Supports 59 (10th ed. 2002) ("AAMR 2002 Manual") (noting that, "[i]n effect, [the SEM] expands the operational definition of mental retardation to 75"); see, e.g. Hardy, 862 F. Supp. 2d at 856-57.

The AAIDD's more recent publications do not make reference to any approximate cutoff score of 75. See AAIDD 2010 Manual; AAIDD 2012 User's Guide. And rightly so. The court can see no particular reason to apply a blanket SEM of 5 points to every kind of IQ test, when

the precise SEMs for the various tests are readily available. See AAIDD 2010 Manual at 27 (significantly subaverage intellectual functioning is "an IQ score that is approximately two standard deviations below the mean, considering the standard error of measurement *for the specific assessment instruments used*" (emphasis added)). Moreover, the courts that have used a range from 5 points below to 5 points above the observed score implicitly assumed without analysis that a 66% confidence interval is appropriate for the interpretation of IQ scores in Atkins and FDPA cases. That may be the case (and indeed, the court concludes below that it *is* the case), but it is certainly not self-evident. Cf. AAIDD 2012 User's Guide at 22 (discussing both 66% and 95% confidence intervals); AAIDD 2010 Manual at 36 (same). Thus, like the AAIDD, the court will depart from the practice of using an approximate upper bound of 75 for a finding of mental retardation. It will apply the precise SEMs applicable to the IQ tests Wilson has taken, and will consider—as a matter of first impression—which percentage confidence interval is appropriate in light of modern clinical literature, the expert opinions, and the nature of these proceedings.

Contrary to Wilson's suggestion (see, e.g., Def. Mem. at 35), the court concludes that a 66% confidence interval—i.e., a range of one SEM below to one SEM above the obtained score—is appropriate in this context for three reasons.

First, the court is aware of no clinical authority (outside of some of the expert opinions in this case) that has expressly advocated for the use of more than one SEM to set the confidence interval. Although the AAIDD's most recent publications mention both a 66% and a 95% percent confidence interval, they do not clarify which one is appropriate. See AAIDD 2010 Manual at 36 ("[A] range of confidence can be established with parameters of at least one standard error of measurement (i.e., . . . 66% probability) or parameters of two standard error[s]

of measurement (i.e., . . . 95% probability)."); AAIDD 2012 User's Guide at 22 (same). And once again, the overwhelming practice prior to the AAIDD's most recent (and inconclusive) statements on the subject has been to set a range from one SEM below to one SEM above the observed IQ score, which corresponds to using a 66% confidence interval. See, e.g., AAMR 2002 Manual at 49; DSM-IV-TR at 41-42; Smith, 790 F. Supp. 2d at 490; Bourgeois, 2011 WL 1930684, at *25; Lewis, 2010 WL 5418901, at *8; Hardy, 762 F. Supp. 2d at 857; Wiley, 668 F. Supp. 2d at 893; Davis, 611 F. Supp. 2d at 475; Thomas, 614 F. Supp. 2d at 1271; see also Atkins, 536 U.S. at 309 n.5. (See also Tr. at 623 (Dr. Olley's testimony).) In other words, it appears that the use of a range from one SEM below to one SEM above the observed score remains the best practice in both the psychological and legal communities. Absent any definitive contradictory statement by the AAIDD, the court has no particular reason to depart from this practice (only reasons to the contrary, as explained below).

Second, and on a similar note, the application of a 95% confidence interval would permit diagnoses of mental retardation well above what has previously been considered the approximate upper bound for a finding of subaverage intellectual functioning. As the AAIDD notes, SEMs for IQ scores typically range from "approximately 3 to 5 points." If the court were to apply a 95% confidence interval to an IQ test with an SEM of 5, then a person could conceivably be diagnosed with mental retardation if his observed IQ score were 80—i.e., two SEMs above 70. So far as the court is aware, no court or clinician has made a finding of mental retardation based on such a high IQ score, and neither the AAIDD nor the APA has ever suggested that such an IQ score would be an indication of significantly subaverage intellectual functioning.

Third, the court considers the use of a 95% percent confidence interval to be particularly inappropriate in the context of an Atkins claim. In the clinical context, there may be good reason

to resolve any doubts in favor of a finding of mental retardation and—as a consequence—the provision of treatment for the disability; in Atkins cases, by contrast, the law places the burden on the defendant to prove his mental retardation by a preponderance of the evidence. Cf. Bourgeois, 2011 WL 1930684, at *26 ("[W]ith the main purpose of most IQ testing being to qualify an individual for needed supports or services, the mental health community has little reason to question the results of IQ testing. Courts, however, endeavor to determine [as a matter of fact] whether a borderline score represents an intelligence capacity above or below the mental-retardation threshold."). Wilson's contention—that the court must ascertain the range of scores in which we are 95% confident that his true score falls—is in serious tension with *his* burden of showing that he is more likely than not mentally retarded.[12]

A final point on confidence intervals. Although the AAIDD maintains that there is no fixed cutoff score for a finding of mental retardation and defines significantly subaverage intellectual functioning as an IQ score of "approximately" 70, these statements are best read simply to mean that it is important to account for measurement error. (See Tr. at 1906 (Dr. Denney: "[M]y judgment is that ['approximately'] means we're talking about standard error.").) See also AAIDD 2010 Manual at 40 ("AAIDD (just as the [APA]) does not intend for a fixed cutoff score to be established for making a diagnosis of ID. Both systems (AAIDD and APA) require clinical judgment regarding *how to interpret possible measurement error*." (emphasis added)); id. at 27 (defining subaverage intellectual functioning as "an IQ score that is

[12] Two of Wilson's experts have advocated for the use of a 99% confidence interval (see Tr. at 102 (Dr. Shapiro's testimony); James IQ Chart (Dkt. 982-2) at 2), a practice that has no basis in any clinical literature of which the court is aware. Dr. Shapiro asserted that when "talking about [a] person's life, we probably should be using the 99th percentile and not having a five percent chance of being off." (Tr. at 103.) The court is certainly aware of the high stakes in an Atkins proceeding, but the fact remains that the burden rests on Wilson to prove that he is more likely than not mentally retarded, a principle that is inconsistent with the overwhelming presumption that Drs. Shapiro and James wish to apply in his favor. See generally Hedges v. Obama, No. 12-CV-331 (KBF), 2012 WL 3999839, at *42 (S.D.N.Y. Sept. 12, 2012) ("A preponderance standard simply asks whether a fact is more likely than not—51 percent likely . . . ." (internal quotation marks omitted)).

approximately two standard deviations below the mean, *considering the standard error of measurement* for the specific assessment instruments used and the instruments' strengths and weaknesses" (emphasis added)). But if the *bottom* of the confidence interval for a well-administered IQ test is above a 70—that is, *after* adjusting for measurement error—this fact would be strongly indicative that the test-taker is not mentally retarded. For example, if a defendant scores a 71 on the latest Wechsler test, his 66% confidence interval would be from 68.4 to 73.6; such a score may or may not (depending on clinical judgment) be deemed indicative of mental retardation. But if a defendant's 66% confidence interval ranges from, say, 71.4 to 76.6 (the result of a score of 74 on that test), then we would be at least 66% confident that his true score is higher than two standard deviations below the mean, strongly suggesting that he is not retarded. This principle is consistent with the approaches of the APA, see DSM-IV-TR at 41 ("[T]here is a measurement error of approximately 5 points in assessing IQ . . . . Thus, it is *possible* to diagnose Mental Retardation in individuals with IQs between 70 and 75 . . . ." (emphasis added)), and those federal courts that have interpreted IQ scores using an SEM of 5 and an upper bound of 75, see, e.g., Thomas, 614 F. Supp. 2d at 1281 ("A court must [ ] consider . . . the standard error of measurement in determining whether a petitioner's IQ score falls within a *range* containing scores that are less than 70."); Smith, 790 F. Supp. 2d at 490; Bourgeois, 2011 WL 1930684, at *25; Hardy, 762 F. Supp. 2d at 857; Blue, 2010 WL 8742423, at *9; Davis, 611 F. Supp. 2d at 475; see also Atkins, 536 U.S. at 309 n.5.

In sum, when interpreting Wilson's IQ scores in light of their inherent potential to be affected by measurement error, the court will: (1) apply the particular SEMs for the different IQ tests Wilson has taken; and (2) use a 66% confidence interval—the range from one SEM below to one SEM above the observed score. An IQ score for which the 66% confidence interval

encompasses scores of 70 or below may be indicative of mental retardation depending on clinical judgment, but an IQ score with a 66% confidence interval *beginning* above 70 will be considered strong evidence against a finding of mental retardation.

2. The Flynn Effect

A second important issue is whether the court should take the so-called "Flynn Effect" into account when considering Wilson's IQ scores. The court concludes that it should.

The Flynn Effect is a phenomenon discussed in a series of widely-cited papers by James R. Flynn. See, e.g., James R. Flynn, Tethering the Elephant: Capital Cases, IQ, and the Flynn Effect, 12 Pysch., Pub. Policy, & L. 170 (2006). Flynn's studies show that in almost all nations in the developed world, there is an upward trend in performance on a given IQ test after the test is normed. See Davis, 611 F. Supp. 2d at 485. In other words, populations as a whole will do better on an IQ test as time passes after the norming of the test. Id. "The cause of this increase is largely unknown, although some speculate that improved socioeconomics, education and even better nutrition have increased the scores, that the test[s] themselves have become more sophisticated, or that perhaps people are simply getting 'smarter.'" Hardy, 762 F. Supp. 2d at 858; see also AAIDD 2010 Manual at 37. The amount of increase varies depending on the test, but it is approximately 3 IQ points per decade, or 0.33 points per year. Davis, 611 F. Supp. 2d at 485-86.

What this means is that over time, the norms for an IQ test become outdated. Id. at 486. Because the population has improved (or has gotten "smarter") since the time the test was normed, the average score for the population would be above 100 if the test were given to a representative sample today. Hardy, 762 F. Supp. 2d at 858. This also means that a score two standard deviations below the average—the approximate score for a finding of mental

retardation—will be higher than 70. Davis, 611 F. Supp. 2d at 486. In other words, although mentally retarded people are getting "smarter" along with the rest of the population, they remain mentally retarded because of their level of intelligence *relative* to the average member of the population (two standard deviations away). For example, someone who receives a score of 72 on a test that was normed a decade ago would be expected, on average, to score a 69 on a newly-normed test. Hardy, 762 F. Supp. 2d at 858. His intellectual functioning is still two standard deviations below the mean in spite of his above-70 score on the outdated test.

"Professionals in the field do not disagree that the [Flynn Effect] exists, but rather, there is professional disagreement regarding whether to adjust an individual's score" to correct for the Flynn Effect. Wiley, 668 F. Supp. 2d at 894; see also Thomas v. Allen, 607 F.3d 749, 753 (11th Cir. 2010) ("The parties in this case agree that the Flynn effect is an empirically proven statistical fact; however, they disagree on the extent to which an individual test subject's IQ score should be adjusted to take into consideration this phenomenon."). Generally, adjustments for the Flynn Effect are made by lowering the individual's IQ score by 0.33 points for each year after the test was normed, before accounting for the SEM. See Lewis, 2010 WL 5418901, at *8; AAIDD 2010 Manual at 37.

In Flynn's view, adjusting IQ scores on the basis of the Flynn Effect is necessary because, "[o]therwise, one person will meet the criterion of mental retardation, and another person will be judged not to have done so, purely because one took a test with current norms and the other took a test with obsolete norms." Flynn, supra, at 176. The AAIDD recommends adjusting for the Flynn Effect for similar reasons. See AAIDD 2010 Manual at 37 (stating that "best practices require recognition of a potential Flynn Effect when older editions of an intelligence test (with corresponding older norms) are used in the assessment or interpretation of

an IQ score," and that "[i]n cases where a test with aging norms is used, a correction for the age of norms is warranted"); AAIDD 2012 User's Guide at 23. Experts in this case from both sides have recognized the existence of the phenomenon. (See, e.g., Tr. at 1959, 1998-99, 2001 (Dr. Denney's testimony); Shapiro Rep. at 7; James Rep. at 11; Woods Rep. at 5.) And among courts, "[w]hile support for the use of the Flynn effect to adjust IQ scores . . . may not be universal, it is widespread." Davis, 611 F. Supp. 2d at 486; see, e.g., Walker v. True, 399 F.3d 315, 322 (4th Cir. 2005) (criticizing a district court for refusing to consider the Flynn Effect and directing the court to consider it on remand); Smith, 790 F. Supp. 2d at 491 ("The Court [ ] finds the Flynn Effect should be applied to the [IQ] scores."); Lewis, 2010 WL 5418901, at *11 ("[T]he court recognizes the Flynn Effect as a best practice for an intellectual disability determination."); Hardy, 762 F. Supp. 2d at 862-63 ("In light of the substantial evidence supporting the existence of the Flynn Effect, the Court concludes that [defendant's] score of 73 should be corrected to take it into account."); Wiley, 668 F. Supp. 2d at 894 (accounting for the Flynn Effect); Davis, 611 F. Supp. 2d at 488 ("[T]he Court finds the defendant's Flynn effect evidence both relevant and persuasive, and will, as it should, consider the Flynn-adjusted scores in its evaluation of the defendant's intellectual functioning."); Thomas, 614 F. Supp. 2d at 1278 ("It [ ] is undisputed that Professor Flynn's recommendation . . . is a generally accepted adjustment."); Green v. Johnson, No. 05-CV-3540, 2006 WL 3746138, at *45 (E.D. Va. Dec. 15, 2006) ("[T]he Flynn Effect is recognized throughout the profession.").[13]

[13] A minority of courts have refused to adjust for the Flynn Effect, or have concluded that a state court's failure to adjust was not an unreasonable application of clearly established federal law for the purposes of adjudicating a writ of habeas corpus pursuant to 28 U.S.C. § 2254. See, e.g., In re Mathis, 483 F.3d 395, 398 n.1 (5th Cir. 2007) ("The Flynn Effect . . . has not been accepted in this Circuit as scientifically valid . . . ."); Candelario-Santana, 2013 WL 101615, at *15 ("[T]he Flynn Effect has no relevance to our inquiry . . . ."); Williams, 2012 WL 4505774, at *34 (because "[t]he Atkins decision does not [ ] mandate that state courts take the . . . Flynn Effect into account[, s]everal Circuit Courts to consider the issue have [ ] determined that a state court's failure to account for [it] is not 'contrary to clearly established federal law' for purposes of § 2254(d)(1)" (collecting cases)); Ledford v. Head, No. 02-CV-1515 (JEC), 2008 WL 754486, at *7 (N.D. Ga. Mar. 19, 2008).

Despite this substantial authority, and without actually contesting the validity of the Flynn Effect, the Government argues that the court should not adjust Wilson's scores based on the Flynn Effect for two reasons.

First, the Government argues that the Flynn Effect was not widely recognized until well after Atkins, and thus its acceptance by the AAIDD and other clinical publications should not be deemed "part of the national consensus against executing the mentally retarded." (Gov't Mem. at 28-29.) The court has already rejected the argument that it is precluded from considering post-Atkins developments in the clinical understanding of mental retardation, and the use of the Flynn Effect is one such development. (See Part II.A.)

Second, the Government argues that the court should not adjust Wilson's scores based on the Flynn Effect because such adjustments are not typically made in standard clinical practice (as opposed to the forensic context). (See Gov't Mem. at 26.) See also Candelario-Santana, 2013 WL 101615, at *15 ("[T]he government's experts could not point to a single instance in their professional experience where they applied, or could recall a colleague's application of, the Flynn Effect."); Ledford v. Head, No. 02-CV-1515 (JEC), 2008 WL 754486, at *7 (N.D. Ga. Mar. 19, 2008) ("There was testimony at the hearing that the Flynn effect is a generally recognized phenomenon, but experts for both petitioner and respondent agreed that it is not used in clinical practice to reduce IQ scores."). "That may be so, but it does not justify ignoring the phenomenon in the face of its unchallenged existence." Thomas, 614 F. Supp. 2d at 1277. The court's duty is to determine as best it can whether Wilson has significantly subaverage intellectual functioning—an IQ approximately two standard deviations below the mean. Attempting to do so without adjusting for indisputably outdated norms seems senseless.

3. The Practice Effect

In addition to the Flynn Effect, the AAIDD recommends considering a phenomenon called the "practice effect." AAIDD 2010 Manual at 35, 38, 102. The court will do so, but not in the manner that Wilson requests.

The AAIDD describes the practice effect as follows:

> The practice effect refers to gains in IQ scores on tests of intelligence that result from a person being retested on the same instrument. . . . For example, the [Wechsler Adult Intelligence Scale-Third Edition] Manual presents data showing the artificial increase in IQ scores when the same instrument is readministered within a short time interval. The [ ] Manual also reports average increases between administrations with intervals of 2 to 12 weeks. For this reason, established clinical practice is to avoid administering the same intelligence test within the same year to the same individual because it will often lead to an overestimate of the examinee's true intelligence.

Id. at 38 (citation and italics omitted); see also AAIDD User's Guide at 23.

The theory behind the practice effect "is that because IQ assessments rely upon novel tasks and instructions to assess ability and performance, an instruction given on a test will be more familiar to the examinee and more quickly implemented on subsequent presentations." Wiley, 668 F. Supp. 2d at 896. The effects are normally greater on performance items than on verbal items. (See James Rep. at 12 ("Performance scales are more susceptible to practice effects because the tasks rely, in part, on the novelty of items and familiarity with those items takes away the novelty, improving scores.").)[14]

Wilson argues that "the Court should apply at least a 5 to 8 point adjustment to the more recent full scale IQ scores similar to the Flynn Effect." (Def. Mem. at 32.) He apparently bases

[14] Contrary to the Government's suggestion (see Gov't Opp'n at 36), the fact that an individual does not exhibit an increase in his IQ scores does not mean that these scores have not been influenced by the practice effect. The practice effect may be offset by other factors tending to diminish a score, such as examiner error, cooperation of the test-taker, or measurement error, which may result in an unchanged overall score. (See Tr. at 1213-14 (Dr. James), 1929-31 (Dr. Denney).)

this argument entirely on the views—relied upon in Dr. James's report (see James Rep. at 12)—of Alan. S. Kaufman, a psychology professor who did not testify at the Atkins hearing (see Def. Mem. at 32). According to Dr. Kaufman:

> Clinicians should understand the average practice effect gains in intelligence scores for children, adolescents, and adults. The expected increase of about 5 to 8 points in global IQ renders any score obtained on a retest as a likely overestimate of the person's true level of functioning—especially if the retest is given within about six months of the original test, or if the person has been administered a Wechsler scale (*any* Wechsler scale) several times in the course of a few years.

Alan. S. Kaufman, Practice Effects, in Robert J. Sternberg, 2 Encyclopedia of Human Intelligence 828 (1994).

As an initial matter, the court is reluctant to place significant reliance upon Dr. Kaufman's views. Because he did not testify at the Atkins hearing, he has not been qualified as an expert in this case, the Government has not had the opportunity to cross-examine him, and the court has not had the opportunity to judge his credibility.

In any event, neither Dr. Kaufman's article nor any other psychological literature provided by Wilson supports the application of a blanket 5 to 8 point adjustment for all IQ retests. Unlike with the Flynn Effect, there does not appear to be an accepted method in the psychological community for adjusting IQ scores to account for the practice effect. (See Tr. at 626 (Dr. Olley's testimony that "the Practice Effect is known to inflate scores, but there's no formula that says if you take this test, this often, this much time in between tests, that it will inflate the score by this number of points").) See also Green, 2006 WL 3746138, at *44 ("There was no consensus among the experts on the degree to which the practice effect would effect a test score, especially for someone with low intellectual functioning."). Contrary to Wilson's suggestion, Dr. Kaufman does not recommend *adjusting* an individual's IQ scores by 5 to 8 points for all retests; he recommends only that "[c]linicians should *understand* the average

practice effect gains." Id. (emphasis added). The AAIDD also does not recommend any particular adjustment of scores to account for the practice effect (as it does with the Flynn Effect); it recommends only that clinicians "avoid administering the same intelligence test within the *same year*." AAIDD 2010 Manual at 38 (emphasis added). Indeed, Wilson's own expert, Dr. Olley, suggested that although IQ scores should be "interpreted" based on the practice effect, they should *not* be adjusted—or "reduced" on this basis. (Tr. at 625-27.) See also Green, 2006 WL 3746138, at *44 (refusing to adjust a petitioner's score to account for the practice effect because "a conclusion that the practice effect increased [petitioner's] test scores by a certain number of points would be purely speculative.").

It is unsurprising that Wilson's proposed adjustment lacks support because the practice effect is heavily dependent upon the length of time between the original test and the retest. Dr. Denney, for example, testified that there is an important difference between a retest within "a short period" and a retest after a period of more than a year, and that practice effects tend to "fall away" after seven years. (Tr. at 1920.) Cf. Kaufman, supra, at 828 (practice effect overestimates a person's intellectual functioning "especially if the retest is given within about six months of the original test, or . . . several times in the course of a few years"). A number of courts have also recognized this principle. See, e.g., Blue, 2010 WL 8742423, at *13 ("[T]he practice effect only applies when there is a short interval between tests. The nine-month period here should have dispelled any lingering effect from the first test."); Garcia Briseno, 2007 WL 998743, at *8 ("[I]n a two-to-twelve week period of retesting, full-scale IQ can be as much as five points higher. For performance IQ scale practice effects will be minimized after between a year to two years."); Green, 2006 WL 3746138, at *44 ("The practice effect refers to an increase in a person's score on an IQ test when it is administered *within a short time* after taking the same or [a] similar test.

. . . [T]he effect is more pronounced the closer in time the tests are given." (emphasis added)); Bowling v. Commonwealth, 377 S.W.3d 529, 539 (Ky. 2012) (the practice effect "refers to only relatively short term periods between tests").

To be sure, experts in this case have suggested that the practice effect *could* occur after even a very long interval between administrations. (See, e.g., Tr. at 1214-1215 (Dr. James: "[T]here's research that shows practice effects can happen after a few weeks, they can happen after a few years, they can happen as much as 13 years on from the original test. . . . They can occur at any interval."); id. at 1920-21 (Dr. Denney's testimony that practice effects could "[p]otentially" affect scores seven years after the first test and, in "rare instances," twelve to thirteen years later).) But no expert has suggested that, after a very long interval, the practice effect would still be expected to increase a person's test scores by an average of 5 to 8 points, or for that matter by any particular number of points. Wilson has provided the court with no persuasive guidance as to the proper adjustment (if any) for IQ tests taken a substantial amount of time after the original test, and the evidence suggests that any such adjustment should be minimal.

To summarize, the court will—as the AAIDD recommends—take into account the practice effect in interpreting Wilson's IQ scores. See AAIDD 2010 Manual at 35, 38, 102. But the court will not apply any particular point adjustment to his IQ scores on the basis of the practice effect, as it finds no basis for doing so in the psychological literature or case law. Cf. Green, 2006 WL 3746138, at *44. In other words, the court will take Dr. Olley's approach and *interpret* Wilson's IQ scores in light of the practice effect without "reducing" his scores on that basis. (Tr. at 626.) When doing so, the court will be mindful that the practice effect diminishes

significantly (although perhaps without disappearing entirely) as the length of time between test administrations increases. See Blue, 2010 WL 8742423, at *13.

4. Raw Data

IQ scores are calculated on the basis of "raw data," including "the testing booklet, the scoring computations, and other relevant data." (James Rep. at 10.) According to Dr. James, "it is essential to be able to review the raw data for each previous [IQ] test administered in order to give that test appropriate weight towards understanding a person's functioning." (Id. at 7.) She argues that raw data is important for three purposes: (1) to determine whether a test has been "completely administered"—that is, administered using all of the proper subtests; (2) to determine whether "scoring and clerical errors" distorted the outcome; and (3) to help account for the practice effect. (Id. at 9.) Thus, Dr. James argues, "the tests which lack raw data should be considered, but given little weight in determining Mr. Wilson's intellectual functioning." (Id. at 9-10.) Indeed, she asserts that "the *only* IQ testing that can be relied upon in this case are tests for which raw data exists." (Id. at 10 (emphasis added).) This approach conflicts with that of the Government's experts, who relied upon all of Wilson's IQ scores, including those lacking in available raw data. (See, e.g., Denney Rep. at 44-45; Mapou Rep. at 20-23.)

The court is cognizant of the risk of errors in IQ administration, and that the availability of raw data makes it possible to check more carefully for (although by no means rule out entirely) such errors. Thus, the unavailability of raw data is a factor to be considered in determining the relative weight to give a particular IQ score. The court disagrees, however, with Dr. James's view of this significance of this factor.

For one thing, errors in IQ test administration are already largely taken into account by the interpretation of IQ tests using the SEM and confidence intervals. (See Part III.A.1.) See

also AAIDD 2010 Manual at 36 ("An IQ score is subject to variability as a function of a number of potential sources of error, including variations in test performance, examiner's behavior, cooperation of the test taker, and other personal and environmental factors. . . . The term standard error of measurement . . . is used to quantify this variability."). The widespread use of the SEM to interpret IQ scores strongly suggests that the psychological community takes as a given that IQ test administration is prone to error, but that this error is not frequent or significant enough to invalidate the scores. Given that IQ tests are typically given by licensed professionals trained in intelligence test administration, the court sees good reason for such a presumption.

Indeed, as noted below (see infra n.21), of the three IQ tests in this case for which raw data was available, only one contained an error that was revealed by an analysis of that data, and this error resulted only in a single-point increase in Wilson's full-scale IQ score—still within the 66% confidence interval around Wilson's observed score for that test (see James Rep. at 8). The court sees little reason to disregard or give substantially diminished weight to certain IQ tests because of the mere *possibility* of these kinds of mistakes.

Perhaps most importantly, it bears repeating that Wilson has the burden of proof on his Atkins claim. He may not satisfy that burden by speculating about the possibility of error in the tests that undermine his Atkins claim and then asking the court to focus primarily on the tests that support his claim, simply because the availability of raw data makes the latter tests somewhat—and, from what the court can tell, not significantly—more reliable.[15]

[15] For this reason, the cases Wilson cites for the importance of raw data (see Def. Mem. at 21-22) are of little help to him. Those cases discounted *low* IQ scores proffered to *prove* the defendant's mental retardation. See, e.g., Pizzuto, 2012 WL 73236, at *14 (defense expert's IQ score was discredited because he "did not record a full scale score and ha[d] since disposed of his raw data"); Smith v. Ryan, No. 98-CV-234 (TUC) (CKJ), 2012 WL 6019055, at *3 (D. Ariz. Dec. 3, 2012) (discounting a low IQ score because the petitioner's expert "acknowledged that the raw data from [that test] was not available"); Ledford, 2008 WL 754486, at *5 (disregarding IQ scores below 70 because it was "impossible to verify the accuracy or reliability of his scores"). To the extent that these cases placed undue reliance on the lack of available raw data, the court respectfully parts company with them for the reasons discussed

In short, the court finds no persuasive support for Dr. James's sweeping theory of raw data either in psychological literature, case law, or logic. The court will take the availability of raw data into account in interpreting Wilson's IQ scores, but its absence for a particular test will not significantly diminish the weight given to that test's results.

5. The Relationship Between Intellectual and Adaptive Functioning

Finally, there is the question of whether the court should take Wilson's adaptive functioning (prong two of the mental retardation definition) into account in determining whether he suffers from significantly subaverage intellectual functioning (prong one). Both parties and their experts argue that the court should do so, while disagreeing as to whether Wilson's adaptive functioning supports or undermines his position on intellectual functioning. (See Def. Mem. at 12-14; Gov't Mem. at 30-32; Tr. at 1225-26 (Dr. James's testimony); id. at 1907 (Dr. Denney's testimony).) The court finds both parties' positions unpersuasive.

"[A]n assessment of adaptive behavior touches on different things than an IQ test." (Tr. at 886 (Dr. Olley's testimony).) "Adaptive functioning refers to how effectively individuals cope with common life demands and how well they meet the standards of personal independence expected of someone in their particular age group, sociocultural background, and community setting." DSM-IV-TR at 42. The APA definition of prong two requires significant deficits in at least two of ten areas: "communication, self-care, home living, social/interpersonal skills, use of community resources, self-direction, functional academic skills, work, leisure, health and safety." Id. at 49. The "AAIDD takes a more holistic approach and treats adaptive behavior as a global characteristic," Hardy, 762 F. Supp. 2d at 879, finding significant limitations in adaptive functioning where a person performs "approximately two standard deviations below the mean of

---

above. But in any event, given the allocation of the burden of proof, the analyses in these cases are more persuasive than Wilson's argument in this case.

either (a) one of the following three types of adaptive behavior: conceptual, social, and practical or (b) an overall score on a standardized measure of conceptual, social, and practical skills," AAIDD 2010 Manual at 27. The differences between the approaches of the APA and the AAIDD have been described as "mostly theoretical" because both "direct clinicians to the same standardized measures of adaptive behavior," such as the Vineland Adaptive Behavior Scales and the Adaptive Behavior Assessment System-Second Addition. Hardy, 762 F. Supp. 2d at 880 (citing DSM-IV-TR at 42; AAMR 2002 Manual at 76-78, 87-90).

The court struggled throughout the Atkins hearing to elicit an explanation from the parties' experts as to exactly how the adaptive functioning prong interacts with the intellectual functioning prong. No expert gave a particularly clear response, but the court's takeaway was that the experts view the mental retardation definition as something of a sliding scale; that is, if the first prong is a close call, the court may turn to the second prong to nudge it one way or the other. (See, e.g., Tr. at 598, 631 (Dr. Olley's testimony that if an IQ score is "substantially" above 70, there would be no need to conduct an adaptive functioning assessment, but that "when IQ scores are close, it's worth while [sic] taking a look at adaptive functioning"); id. at 1470 (Dr. James: "[W]hen I take a look at these [IQ] scores . . . , they're all low enough for me to . . . look at adaptive functioning . . . ."); id. at 1907 (Defense counsel: [I]f there is some ambiguity about the [IQ] score you can look to [adaptive behavior] deficits to maybe clear up the ambiguity? Dr. Denney: Yes. . . . In extreme cases, no, I think its [sic] probably irrelevant. But in close types of situations I think that that can be of assistance.").)

The court finds this approach fundamentally incompatible with the principle that mental retardation involves three "indispensable" *prerequisites* as opposed to the kind of sliding scale the parties appear to envision. Blue, 2010 WL 8742423, at *3; see also Hall v. State, No.

10-CV-1335, 2012 WL 6619321, at *5 (Fla. Dec. 20, 2012) ("[B]ecause a defendant must establish all three elements of [a mental retardation] claim, the failure to establish any one element will end the inquiry."). The court also finds no convincing reason to limit the sliding scale principle to cases where the intellectual functioning prong is a "close call," as the parties' experts suggested. (See, e.g., Tr. at 598, 631, 1470, 1907.) Under a sliding scale approach, it would seem that a *low* enough IQ could eliminate *any* necessity of finding significant deficits in adaptive functioning. That of course does not represent the accepted clinical standard.

In any event, even assuming that it is proper for *psychologists* to use a holistic approach when interpreting IQ scores in light of their clinical judgment, this does not mean that a *court* should meld the two prongs together when making a legal determination of who is ineligible for the death penalty. For example, Dr. James may reasonably have taken into account Wilson's adaptive functioning when interpreting his IQ scores (although tellingly, she did not do so). For the purposes of the court's analysis, however, where a *legal* test contains multiple necessary prerequisites, a greater showing of one prong cannot overcome a deficient showing in the other, even if the latter is a "close call." See, e.g, Growden v. Ed Bowlin & Assocs., Inc., 733 F.2d 1149, 1150-51 (5th Cir. 1984) ("Th[e] constitutional test [for personal jurisdiction] is two-pronged; the fairness prong cannot compensate for or overcome the requirement of some minimum contacts with the forum state."); Pfizer Inc. v. Teva Pharm. USA, Inc., 820 F. Supp. 2d 751, 759 (E.D. Va. 2011) (where claim contained multiple necessary elements, "no sliding scale [could] be used to compensate one element's weakness with the other element's strength"); cf. Atkins, 536 U.S. at 318 (mental retardation definition requires "*not only* subaverage intellectual functioning, *but also* significant limitations in adaptive skills . . . that became manifest before

age 18 (emphases added)). Because the law is clear that mental retardation contains three necessary elements, the court must determine if these elements are independently satisfied.

For these reasons, the court will not take Wilson's adaptive functioning into account in determining whether he has significantly subaverage intellectual functioning. The court turns now to Wilson himself.

B. Wilson's Intellectual Functioning

1. IQ Test Scores

During his life, Wilson has been administered nine IQ tests—once with the Wechsler Intelligence Scale for Children-Revised ("WISC-R"), five times with the Wechsler Intelligence for Children-Third Edition ("WISC-III"), twice with the Wechsler Adult Intelligence Scale-Third Edition ("WAIS-III"), and once with the Wechsler Adult Intelligence Scale-Fourth Edition ("WAIS-IV"). (See James IQ Charts (Dkt. 982-2).) The chart below displays Wilson's test results. From left to right, it displays: the date the test was given; Wilson's age (years/months);[16] the last name of the test administrator; the test edition; the verbal IQ ("VIQ"), performance IQ ("PIQ"), and full scale IQ ("FSIQ") scores Wilson obtained; the FSIQ after applying an adjustment for the Flynn Effect of 0.33 points per year since the test was normed[17] (see Part III.A.2); the SEM for the test associated with the age of the examinee;[18] and the 66%

[16] Wilson was born on May 6, 1982. (Mapou Rep. at 1.) He committed the murders on March 10, 2003. Whitten, 610 F.3d at 173.

[17] According to Dr. James, the WISC-R was normed in 1972; the WISC-III in 1989; the WAIS-III in 1995; and the WAIS-IV in 2006. (James IQ Charts.) The Government disputes the norming dates of certain IQ tests, arguing that the date reported in the Wechsler manual "as to when the normative data was completed should be the operative date for the Flynn effect, and not a date that Flynn cited in an article that appears to be the year that the norming data was first collected." (Gov't Mem. at 32 n.15.) This issue was not fully sorted out at the Atkins hearing, and the court need not resolve it because of the result reached in this case. The court thus assumes without deciding that Flynn's (and Dr. James's) dates are appropriate for calculating the Flynn Effect.

[18] Dr. James testified that it is more precise to use the average SEM of a test for those of the same approximate age of the examinee than it is to use the average SEM for the full norming population. (See, e.g., Tr. at 1225.) No expert attempted to dispute this approach, and the court accordingly adopts it.

and 95% confidence intervals ("CI") around the Flynn-adjusted scores (see Part III.A.1).[19]

DATE	AGE	EXAMINER	TEST	VIQ	PIQ	FSIQ	FLYNN FSIQ	SEM	66% CI	95% CI
1/06/89	6/8	Abramson	WISC-R	81	90	84	78.39	3.41	74.98 to 81.80	71.57 to 85.21
12/11/91	9/7	Drezner	WISC-III	79	81	78	77.34	3.35	73.99 to 80.69	70.64 to 84.04
10/27/93	11/6	Aranoff	WISC-III	72	90[20]	78	76.68	3.35	73.33 to 80.03	69.98 to 83.38
12/05/94	12/7	Nagler	WISC-III	65[21]	80	70	68.35	3.00	65.35 to 71.35	62.35 to 74.35
4/24/97	14/11	Frank[22]	WISC-III							
4/25/98	15/11	Giglio	WISC-III	70	95	80	77.03	2.60	74.43 to 79.63	71.83 to 82.23
1/07/00	17/8	Popp	WAIS-III	78	92	84	82.35	2.58	79.77 to 84.93	77.19 to 87.51
10/17/03	21/5	Drob	WAIS-III	71	85	76	73.36	2.37	70.99 to 75.73	68.62 to 78.10
6/28/12	30/1	Denney	WAIS-IV	80	92	80	78.02	2.12	75.90 to 80.14	73.78 to 82.26

[19] Although the court concluded in Part III.A.1 that the use of the 66% confidence interval is appropriate for Atkins cases, it will provide the 95% confidence interval as well for illustrative purposes.

[20] Dr. Aranoff reported inconsistent PIQ scores; she twice reported it as 90 and once as 93. (See James IQ Charts.) The court will use the lower of these two scores, which yields an FSIQ of 78. (See id.) If the higher of the two scores were used, the FSIQ would be 80 (the FSIQ Dr. Aranoff reported throughout her report). (See id.)

[21] Dr. Nagler reported a VIQ of 66 and an FSIQ of 71, but her raw data revealed an arithmetic error in the calculation of the VIQ. (See James Rep. at 8; James IQ Charts.) The Government acknowledges that this error occurred (see Gov't Mem. at 6 n.4), and so the court uses the corrected score, which drops the FSIQ to 70.

[22] Dr. Frank administered only five subtests; his testing did not produce an IQ score. (James Rep. at 11; see also Part III.B.3.)

2. Preliminary Analysis of the IQ Test Scores

The court will carefully consider below the clinical judgment of Wilson's test administrators and the experts in this case regarding how the above IQ scores and the full record should be interpreted. It first, however, conducts the following preliminary analysis of those scores. The court's initial finding is that Wilson's IQ scores appear simply too high to qualify him under the definition of significantly subaverage intellectual functioning.

After adjusting for the Flynn Effect, seven of Wilson's eight IQ scores are at least 3 points above 70—the benchmark for significantly subaverage intellectual functioning. See AAIDD 2010 Manual at 27; DSM-IV-TR at 49. Wilson's average Flynn-adjusted FSIQ for these tests is 76.44, cf. Davis, 611 F. Supp. 2d at 489 (averaging the defendant's IQ scores), almost a point and a half above the score that many courts have considered to be the upper bound for a finding of mental retardation (see Part III.A.1). The median of his Flynn-adjusted FSIQ scores is 77.19, over two points above the conventional upper bound.

Most importantly, the *bottom end* of the 66% confidence interval is above 70 for seven of his eight scores. The median of the bottom ends of the 66% confidence intervals is 74.21, and the average is 73.59. In other words, on average, Wilson's test scores permit us to say with 66% confidence that his true score lies more than 3.5 points above 70. Indeed, although the court has concluded that the use of a 66% percent confidence interval is appropriate for Atkins cases, even the bottom end of the 95% confidence interval is above 70 for five of his eight scores, the median of the bottoms ends is 71.11, and their average is 70.75.[23]

[23] Wilson argues that his IQ tests are indicative of mental retardation because "all of [his] Flynn- corrected IQ scores except one had overlapping [95%] confidence interval bands that included scores below 75." (Mem. at 48.) Dr. James suggested the same thing offhand at the Atkins hearing. (See Tr. at 1224-25.) This argument misunderstands the purpose of confidence intervals. An IQ score is potentially indicative of mentally retardation if its confidence band includes scores below *70* (two standard deviations below the mean), not 75. See, e.g., Thomas, 614 F. Supp. 2d at 1281 ("A court must [ ] consider . . . the standard error of measurement in determining whether a petitioner's IQ score falls within a range containing scores that are less than 70." (emphasis omitted)). Wilson's

The court makes two additional points about the confidence intervals in this case. First, Wilson's true IQ score is at least as likely to be one SEM *above* his observed score as one SEM below. See Ledford, 2008 WL 754486, at *8. In fact, both Dr. Denney and Dr. James suggested that an examinee's true IQ score is more likely on the higher end of the confidence interval than the lower end, because scores on a bell curve tend to gravitate toward the mean of 100. (See Tr. at 1457, 1949.) The median and average of the upper ends of Wilson's 66% confidence intervals are 80.09 and 79.29, respectively, and the median and average of the upper ends of his 95% confidence intervals are 82.82 and 82.14, respectively—all clearly well above the benchmark for mental retardation. Second, although the court has taken measurement error into account, "there is evidence that measurement error is more of a factor when only one IQ test is given," and is "'much reduced' when more than one IQ test is given and the scores corroborate each other." Ledford, 2008 WL 754486, at *8 (quoting Flynn, supra, at 186). As indicated above, all but one of Wilson's Flynn-adjusted scores fall at least three points above 70, suggesting further that his true IQ score lies in that area.

Next, as discussed above in Part III.A, the court interprets Wilson's IQ tests in light of (1) the possibility of a practice effect in his later tests, and (2) the unavailability of raw data for five of the eight scored tests.[24]

Several considerations diminish the importance of the practice effect in Wilson's case. First, Wilson was never administered two intelligence tests within same year—the procedure that the AAIDD admonishes against. See AAIDD 2010 Manual at 23 ("[E]stablished clinical

---

argument essentially piles three SEMs on top of each other—a blanket SEM of 5 to raise the benchmark for mental retardation up to 75, and then two more SEMs to create a 95% confidence interval *around 75*. There is no basis for this practice.

[24] Wilson attempts to discount a number of his IQ scores for reasons specific to the particular test; the court addresses these arguments in Part III.B.3.

practice is to avoid administering the same intelligence test within the same year to the same individual . . . ."); see also Blue, 2010 WL 8742423, at *13 ("The nine-month period [between test administrations] should have dispelled any lingering effect from the first test."); Garcia Briseno, 2007 WL 998743, at *8 ("[P]ractice effects will be minimized after between a year to two years."); Green, 2006 WL 3746138, at *44; Bowling, 377 S.W.3d at 539. Second, Wilson's later FSIQ scores were substantially consistent with the scores he obtained early in his life, suggesting that the practice effect may be present but, if so, not particularly significant: his first three Flynn-adjusted scores were 78.39, 77.34, and 76.68 (an average of 77.47), and his last three were 82.35, 73.36, and 78.02 (an average of 77.91).[25] Third, Wilson's most recent IQ test in 2012, which resulted in a Flynn-adjusted FSIQ of 78.02 and a 66% confidence interval of 75.90 to 80.14, was administered more than eight and a half years after his previous test in 2003, and thus the influence of the practice effect on that test was likely minimal if not nonexistent. (See Tr. at 1920 (Dr. Denney's testimony that practice effects tend to "fall away" after seven years).) In any event, even assuming the practice effect had some effect on Wilson's later test scores—a possibility the court does not find particularly compelling, for the reasons just stated—his scores are high enough that the effect does not change the court's conclusion as to the direction the scores likely point.

The unavailability of raw data underlying five of Wilson's scored IQ tests—all except those administered by Drs. Nagler, Drob, and Denney (see Tr. at 1201)—also does not change the court's analysis. Apart from the reasons discussed in Part III.A.4, two factors specific to this case diminish the importance of raw data. First, the scores that Wilson obtained on the tests for which raw data is unavailable are largely consistent with each other and with those for which raw

[25] His PIQ scores for these tests, which are more likely to be inflated by the practice effect than his verbal scores (see James Rep. at 12) were relatively consistent as well—90, 81, and 90 for the first three, and 92, 85, and 82 for the last three.

data is available, corroborating the validity of the former scores. Second, two of the three scores for which raw data is available are not indicative of significantly subaverage intellectual functioning—the bottom ends of the 66% confidence intervals on Wilson's 2003 and 2012 tests are 70.99 and 75.90, respectively. The one score Wilson obtained that is indicative of mental retardation—on Dr. Nagler's test in 1994—appears to be an outlier.

The bottom line in the court's view is that, even after taking into account the various possibilities for error, see AAIDD 2010 Manual at 27, Wilson's tests strongly suggest that his true IQ score is more likely than not above 70. That is a compelling indication that he does not suffer from significantly subaverage intellectual functioning.

3. Clinical Judgment of the IQ Test Administrators

The court does not rest on its own analysis of Wilson's IQ scores; those scores (and the entirety of the record) must be interpreted in light of "clinical judgment." AAIDD 2010 Manual at 35. The clinicians best situated to interpret Wilson's IQ tests are the individuals who actually administered the tests. (See Tr. at 1241-42 (Dr. James's agreement that "the best person to assess . . . what the intelligence is of an individual, is a person administering that test," and that "there's nothing like actually seeing how [the examinee] answers a question" on an IQ test).) The observations of these clinicians reveal two important points: (1) not one of the clinicians who administered an IQ test to Wilson concluded at the time that he suffered from mental retardation; and (2) most of the test administrators believed that Wilson's observed IQ scores represented an *under*estimate of his true intelligence. Cf. Taylor, 498 F.3d at 307 (crediting a test administrator's testimony that the petitioner "was capable of performing better than a 75," and that the petitioner "was not diagnosed as mentally retarded as a result of the [ ] test");

Bourgeois, 2011 WL 1930684, at *29 (finding "highly credible" an expert's "testimony that [petitioner] underperformed on his testing"). The court reviews the clinical evaluations here.

Wilson's first IQ test was administered by Richard Abramson, Ph.D, when Wilson was six years old, after he was admitted to the psychiatric center of Elmhurst Hospital. (See Denney Rep. at 6.) Wilson scored an FSIQ of 84 (Flynn-adjusted 78.39) on the WISC-R, leading Dr. Abramson to conclude that Wilson "was functioning in the low average range." (Patterson Rep. at 9.[26]) Dr. Abramson noted that Wilson was "functioning below his potential intellectually" because "emotional concerns interfered with his academic and social functioning."[27] (Id.)

In 1991, Wilson was administered the WISC-III by Carla Drezner—a psychologist in Wilson's school system—and scored a 78 (Flynn-adjusted 77.34). (Id.) She noted that Wilson's "borderline I.Q. of 78 (WISC-III) appears depressed as a function of emotional and cultural factors" and that his "true cognitive ability appears to be low average-average." (Id.) Ms. Drezner also testified credibly at the Atkins hearing. When asked why she believed Wilson's true intellect was higher than his IQ score suggested, she testified that he scored in the average to

[26] The notes completed by Wilson's IQ test administrators are set forth in the various expert reports in this case. Neither party disputes any of the experts' characterizations of those notes.

[27] Dr. James testified that a test taken at age six is not a good predictor of someone's IQ as a young adult or adult, primarily because children of that age have insufficient development in the frontal lobes of their brains. (See Tr. at 1206-07.) Dr. Mapou also recognized that "intelligence at a young age does not predict well intelligence in older age" (id. at 2112), and Dr. Denney testified that there were "some weaknesses" in the first score given Wilson's age (id. at 1960). The court takes as a given that Wilson's age at his first IQ test may make that test a less accurate predictor of his intelligence, but notes that this test in one sense is *more* reliable than Wilson's later IQ tests because it could not have been influenced by the practice effect. (See Denney Rep. at 44 (Dr. Abramson's "evaluation is striking because it is the first time Mr. Wilson was exposed to any type of intellectual assessment; therefore, it is free of possible retest effects."); Mapou Rep. at 23 ("The best measure of Mr. Wilson's intellect may have been the first evaluation in 1989, because subsequent evaluations were influenced by . . . possible practice effects.").) The court has little basis for speculating as which one of these factors—Wilson's age or the absence of the practice effect—is more significant, and does not see a need to resolve the issue because, as discussed above, Wilson's score on his first test was substantially in line with his later test scores.

low average range on six of his subtests, and that she believed that his IQ was being pulled down by the "comprehension" and "information" subtests.[28] (Tr. at 1698.)

Wilson's third test was administered in 1993 by Senior Psychologist Ellen Aranoff, Ph.D, at Elmhurst Hospital. (Patterson Rep. at 9.) Wilson obtained an FSIQ of 78 (Flynn-adjusted 76.68) on the WISC-III. (See James IQ Charts; supra n.20.) Dr. Aranoff noted that Wilson's responses to testing were inconsistent: he was cooperative at first, but "on later occasions his anxiety, irritability and preoccupation with personal problems and issues interfered with his ability to respond to presented test questions," and "[i]n these instances, his involvement in examination content was inadequate and detrimental to performing optimally on psychological tests." (Denney Rep. at 9.) Thus, she concluded, Wilson's "present results d[id] not constitute the best estimates of [his] cognitive abilities." (Id.) She further noted the significant difference between Wilson's VIQ and PIQ (the latter of which was in the average range), and explained that "language deficits" may have caused his low functioning on the verbal subtests. (Patterson Rep. at 9.) She opined that he "may have average overall intellectual potential." (Denney Rep. at 9.)

Wilson's fourth IQ test was administered in 1994 by school psychologist Lauren Nagler, Ph.D, as part of a school system triennial evaluation. (See Denney Rep. at 10.) Wilson achieved his lowest score on this test: an FSIQ of 70 (Flynn-adjusted 68.35). (See James IQ Charts; supra n.21.) In addition to its inconsistency with Wilson's other IQ test scores, there is reason to give

[28] Wilson argues that the court should discount Ms. Drezner's test because she gave Wilson the "mazes" subtest instead of the "object assembly" subtest, even though "the latter [is] a core subtest for the performance IQ" portion of the WISC-III. (Def. Mem. at 42.) Ms. Drezner testified that she made this substitution because the mazes subtest was shorter and children liked it more. (See Tr. at 1696-97.) According to Dr. James, this was a departure from standard procedure; the tests measure different skills, and the manual for the WISC-III permits a substitution only when the object assembly subtest is spoiled. (Id. at 1322-33.) Dr. James gave no explanation, however, as to the degree that Ms. Drezner's substitution would have impacted Wilson's IQ score. The court doubts that this impact would have been especially significant given that, according to Dr. James's own testimony, the WISC-III manual contemplates the use of the mazes test as an alternative to the object assembly test in certain circumstances. In other words, the court has no reason to conclude that Ms. Drezner's substitution caused more than a minor inflation to Wilson's score that would be contemplated by the SEM, let alone that this score is invalid.

diminished weight to this score because of Wilson's observed behavior during the exam. Dr. Nagler wrote that Wilson was "resistant and confrontational throughout"; that he "squirmed in place, put his fingers in his mouth, and yawned continuously"; that "[h]e blurted out questions and generally utilized a careless, impulsive approach"; and that when items became "somewhat difficult he became frustrated and gave up." (Denney Rep. at 10.) See also Bourgeois, 2011 WL 1930684, at *27 ("Several Fifth Circuit cases have refused to credit IQ scores when the evidence suggested that an inmate had . . . not put forth his best effort." (collecting cases)). Although Dr. Nagler did not indicate that the results of her test administration were invalid as a result of Wilson's attitude, she did find that, "because of complicating emotional, social and cultural factors, [he] ha[d] not achieved at a level commensurate with ability." (Denney Rep. at 10.) Dr. Nagler concluded that Wilson was "functioning in the borderline-low average range of intelligence" but that there was "evidence of average ability." (Id.)

The next IQ test given to Wilson was in 1997 by Mitchell Frank, Psy.D, as part of a family court proceeding. (Id. at 12.) Dr. Frank administered four verbal subtests and one performance subtest from the WISC-III—not enough to calculate an IQ score—and did not disclose Wilson's scores on the subtests. (Id.) He did note, however, that Wilson's scores on the verbal subtests were "consistent with the mildly deficient range of cognitive abilities," and that his score on the performance subtest "indicated [an] average level of abilities." (Id.) Dr. Frank estimate that Wilson's intelligence was "in the borderline range." (Id. at 13.)

In 1998, Wilson was administered the WISC-III by John Giglio at the Brookwood juvenile detention center. (Id. at 14.) Wilson obtained an FSIQ of 80 (Flynn-adjusted 77.03), but Mr. Giglio considered the FSIQ invalid due to the large difference between Wilson's VIQ

(70) and PIQ (95). (Id.) He found that Wilson's "pattern of scores indicate[d] a Learning Disability in language abilities."[29] (Id.)

A few months before Wilson's eighteenth birthday—which Dr. Denney described as "a critical point in the retrospective analysis of potential intellectual disability" (Denney Rep. at 44)—he was administered a seventh IQ test by Arthur Popp, Ph.D, for the purposes of placement at the Far Rockaway School (id. at 16). Wilson scored an FSIQ of 84 (Flynn-adjusted 82.35) on the WAIS-III, which Dr. Popp found to be in the "low average range." (Id.) Dr. Popp further noted the "clearly average outcomes for tasks concerned with abstract reasoning and practical and social knowledge, suggesting the capability to operate verbally" (id.), and that Wilson had the "potential to function in the mainstream" (Patterson Rep. at 13).[30]

---

[29] There was a great deal of discussion and disagreement at the Atkins hearing as to whether (and if so, how often) a learning disability might coexist with mental retardation (so-called "comorbid conditions"). (See, e.g., Tr. at 342 (Dr. Shapiro's testimony that "you can have intellectual disability and a learning disability at the same time"); id. at 621 (Dr. Olley's agreement that it was "extremely rare" for a learning disability to coexist with mental retardation); id. at 1425 (Dr. James's testimony that "[i]f one is diagnosed with mental retardation [one] can also be diagnosed with a learning disability," and that "the learning disability in someone with an intellectual disability is not unexpected because they have broader deficits across other domains"); id. at 1560-61 (Dr. Woods's testimony that it would be "rare" for someone to have a learning disorder and mental retardation at the same time).) The court need not resolve this issue because of the outcome of its independent analysis of Wilson's alleged mental retardation, and therefore assumes without deciding that mental retardation can coexist with a learning disability.

[30] Dr. Popp administered only four of the five performance subtests—all but the "picture arrangement" subtest— and he therefore "prorated" Wilson's PIQ and FSIQ scores; that is, he estimated these scores based on the four subtests that he administered. (Mapou Rep. at 22.) The WAIS-III Administration and Scoring Manual (1997) ("WAIS-III Manual") states: "On occasion, a subtest may be spoiled or impossible to administer. In these cases, it is recommended that an alternative subtest be administered in its place. If an alternative subtest is not available, you can prorate the IQ scores . . . ." WAIS-III Manual at 38; see also id. at 59 ("To prorate the examinee's scores on the Performance subtests, multiply the sum of the scaled scores by 1.25, round to the nearest whole number, and enter the result . . . ."). Wilson argues—relying upon the testimony of Drs. Shapiro and James—that Dr. Popp's score should be given less weight because it was prorated, particularly because Wilson later obtained a low score on the picture arrangement subtest when it was administered to him in 2003. (Def. Mem. at 26-27 (citing Tr. at 117-18, 410-11, 1219-21).) The court's response to this argument is similar to its reasoning regarding Ms. Drezner's substitution of a performance subtest (see supra n.28): given that the WAIS-III permits prorating in certain circumstances, the court doubts that Dr. Popp's omission of a single subtest (perhaps with good reason to do so) and prorating of the scores would impact the FSIQ too significantly or to a degree not accounted for by the SEM, let alone invalidate it. And given how high Wilson scored, it is even more doubtful that the administration of the picture arrangement subtest would have caused his FSIQ to drop into the range that would be indicative of mental retardation.

In 2003, just over six months after the crimes at issue in this case, Wilson was evaluated by Sanford Drob, Ph.D, at the request of his defense attorneys. (See Denney Rep. at 18; James Rep. at 7.) Dr. Drob administered the WAIS-III and a full neuropsychological test battery. (James Rep. at 8.) Wilson obtained an FSIQ of 76 (Flynn-adjusted 73.36), his second-lowest score. Dr. Drob wrote that the deficits in Wilson's "verbal abilities [were] greater than [his] deficits in his non-verbal abilities," and that his "normal range and higher scores on two non-verbal subtests suggests a higher intellectual potential." (Patterson Rep. at 13; see also Tr. at 1896 (Dr. Denney: "In seven out of the eight cases, there was a significant split [between verbal and performance scores] and that pattern is, frankly, not consistent with intellectual disability in my opinion.").) Dr. Drob "rule[d] out the likelihood of mental retardation." (Tr. at 1019.) He also testified that although he did not adjust Wilson's score to account for the Flynn Effect, such an adjustment would not have changed his conclusion. (Id. at 992.)

Wilson's final IQ test was administered by Dr. Denney in connection with this Atkins proceeding. Dr. Denney administered Wilson the WAIS-IV—the most current edition of the Wechsler exams—and Wilson obtained an FSIQ of 80 (Flynn-adjusted 78.02, with a 66% confidence interval of 75.90 to 80.14), which Dr. Denney noted was "in the low average to borderline range of functioning." (Denney Rep. at 41.) Dr. Denney concluded, in light of the IQ test he administered and "the entirety of the record," that Wilson's "true intellectual capacity was likely in the lower portion of the low average range prior to his 18th birthday." (Id. at 45.) The court finds Dr. Denney's IQ test administration particularly compelling because: (1) the WAIS-IV is indisputably the best current method of measuring IQ (see Tr. at 659 (Dr. Olley's agreement that the WAIS-IV is currently the "gold standard of IQ tests"); see also James Rep. at 3 ("Each new generation of IQ test instruments . . . has developed a better estimate of [human

intelligence] and its components.”));[31] (2) the raw data for Dr. Denney’s test is indisputably complete, and revealed no scoring errors when reviewed by Dr. James (see Tr. at 1204; Def. Mem. at 46); and (3) Dr. Denney’s test was administered more than eight and a half years after Wilson’s previous test in 2003, rendering minimal the potential influence of the practice effect (see Tr. at 1920 (Dr. Denney’s testimony that retest effects tend to “fall away after seven years”); see also Blue, 2010 WL 8742423, at *13; Garcia Briseno, 2007 WL 998743, at *8; Green, 2006 WL 3746138, at *44; Bowling, 377 S.W.3d at 539.

In sum, the court’s own analysis of Wilson’s IQ scores is supported by the observations of the clinicians who administered his tests. None of these clinicians believed that Wilson suffered from mental retardation. And most of them believed that Wilson’s scores underestimated his intellectual functioning.

4. Clinical Judgment of the Parties’ Experts

Having concluded that Wilson’s IQ tests and the opinions of the clinicians who administered those tests favor a finding that he is not mentally retarded, the court turns next to the parties’ principal medical experts in this case.

Each of Wilson’s experts opined that he suffers from significantly subaverage intellectual functioning (see James Rep. at 1-2; Olley Rep. at 8; Shapiro Rep. at 2; Woods Rep. at 21), and each of the Government’s experts opined that he does not (see Denney Rep. at 45, 48; Patterson Rep. at 18; Mapou Rep. at 23, 31). However, only three experts—Dr. James, Dr. Denney, and

[31] Strangely, Wilson suggests that Dr. Denney’s test should be *discounted* because he used the WAIS-IV. (See Def. Mem. at 47.) Wilson notes that because the WAIS-IV eliminated the speed tests that were part of the WAIS-III, and because Wilson “‘tends to be very slow,’ [ ] he would be expected to do better on the WAIS-IV where speed no longer mattered.” (Id. (quoting Tr.at 119 (Dr. Shapiro’s testimony).) But the creators of the WAIS-IV presumably eliminated the “speed tests” because they believed that those tests were a less reliable measure of intelligence than the replacement tests. See James Rep. at 3 (“Each new generation of IQ test instruments . . . has developed a better estimate of [human intelligence] and its components.”).) That may be reason to give less weight to Wilson’s low scores on the speed tests in earlier administrations, not to discount his score on the WAIS-IV.

Dr. Mapou—conducted a robust analysis of Wilson's intellectual (as opposed to adaptive) functioning. (Cf. Olley Rep. at 8 (noting that his "evaluation focused on . . . adaptive behavior" but that he had reviewed Dr. James's report and concurred in her opinion); Shapiro Rep. at 2 (noting that he had not personally examined Wilson but had relied "on the work of other members of the evaluating team," including Dr. James); Woods Rep. at 21 (providing a one-sentence opinion on the intellectual functioning prong, namely that Wilson satisfied this requirement "[f]or the reasons stated in the reports of Dr. James, Dr. Shapiro, and Dr. Olley"); Patterson Rep. at 18 (providing a one-paragraph analysis of mental retardation after summarizing Wilson's medical record and the results of an in-person examination).) Moreover, the opinions of Drs. Denney and Mapou have been relied upon above and are consistent with the court's analysis thus far, so the court does not discuss their reports in any significant additional detail.

The court therefore focuses in on Dr. James, the only expert in this case who both: (1) performed a substantial analysis of Wilson's intellectual functioning; and (2) concluded that he satisfies the intellectual functioning prong. Dr. James is undoubtedly a well-respected expert in her field, and provided thoughtful and well-reasoned testimony in this case. Nevertheless, in addition to the inconsistency of her opinions with the evidence discussed above, the court finds a number of reasons to question her judgment.

Foremost among these reasons is the fact that Dr. James (unlike Dr. Denney) did not administer an IQ test on Wilson. Given that IQ tests are—as Dr. James acknowledged (see Tr. at 1241)—currently the best method of measuring intelligence, this omission significantly undermines her opinion. See AAIDD 2010 Manual ("Although far from perfect, intellectual functioning is currently best represented by IQ scores when they are obtained from appropriate, standardized and individually administered assessment instruments."); Hardy, 762 F. Supp. 2d at

875 ("Both the APA and AAMR/AAIDD indicate that a diagnosis of mental retardation should be made based on IQ test results where it is possible to perform such a test."); Davis, 611 F. Supp. 2d at 507 (giving less weight to a government expert's opinion in part because he "failed to administer an IQ test" to the defendant). Dr. James's failure to administer an IQ test is particularly troubling because she had the opportunity to administer the WAIS-IV—the "gold standard" of IQ tests (Tr. at 659) and one that Wilson had never been previously administered. Dr. James did administer a substantial battery of other kinds of neuropsychological testing (see James Rep. at 13-17), but the results of this testing would have been more compelling as a complement to an IQ test.

Indeed, the court can find no reasonable basis for Dr. James's failure to administer an IQ test. Wilson argues that this failure "should not be held against" him because: (1) the defense "could not legitimately administer yet another Wechsler instrument and then turn around and credibly talk about practice effects"; and (2) "the issue at hand was intellectual disability at the time of the crime, not years later." (Def. Mem. at 34.) Neither of these arguments is convincing.

Regarding the practice effect, the court notes once again that it had been over nine years since Wilson's previous IQ test administration, and so the potential influence of the practice effect was likely not particularly significant (see Tr. at 1920 (Dr. Denney's testimony)); see also Blue, 2010 WL 8742423, at *13; Garcia Briseno, 2007 WL 998743, at *8; Green, 2006 WL 3746138, at *44; Bowling, 377 S.W.3d at 539, let alone significant enough to render any additional IQ testing *worthless*. Moreover, Dr. James agreed that "the best person to assess" a person's intelligence is the "person administering the [IQ] test." (Tr. at 1241.) By administering an IQ test, Dr. James could have placed herself in a far better position to observe whether Wilson's scores were in fact being inflated by the practice effect and, if so, to what degree.

The court is also unconvinced by Wilson's argument as to the time that Dr. James would have administered an IQ test. (See Def. Mem. at 34.) It is true that, to be exempt from the death penalty, Wilson must show that he was mentally retarded at the time of the crime. See Holladay, 555 F.3d at 1353; Hardy, 762 F. Supp. 2d at 881. But it is also true that IQ "is a relatively stable, immutable trait," and so "[a] person's IQ tested after the developmental period is, absent intervening trauma or injury, likely to be quite close to the IQ that would have been obtained had the person been tested" earlier."[32] Hardy, 762 F. Supp. 2d at 881-882 (citing AAMR 2002 Manual at 51-59). Indeed, if Wilson's argument were correct, almost *all* of his IQ tests would be irrelevant to this case, including Dr. Nagler's in 1994—over eight years before the crime—upon which he places principal reliance.

In short, the court is troubled that Wilson's primary expert on intellectual functioning did not make use of the best current method of assessing intelligence. Surely no harm would have resulted from doing so.

A number of other factors also lead the court to give less credit to Dr. James's opinion. Her opinions were at times inconsistent and at other times unduly selective or dismissive in emphasizing certain pieces of evidence over others. For example:

As noted above in Part III.A.5, Dr. James testified repeatedly that an evaluation of a person's intellectual functioning should take into account his adaptive functioning (see, e.g., Tr. at 1225), and yet she herself did not address Wilson's adaptive functioning in the analysis she conducted in her expert report, other than simply to express agreement with the opinions of

[32] As discussed above, the Flynn Effect and the practice effect at times cause variation in a person's scores over time, but these phenomena are dependent on, respectively, the age of the test and the number of times a person has been administered a similar test, not on the age of the person at the time of administration. In other words, all else being equal, a person's IQ score is no more likely to be influenced by these effects at age thirty than at age eighteen.

Wilson's other experts (see James Rep. at 2). In other words, Dr. James's evaluation was incomplete under her own standards.

Dr. James was also too quick to dismiss the opinions of most of Wilson's test administrators that his IQ scores were an underestimate of his true intelligence. (See Part III.B.3.) While recognizing that these clinicians were in the best position to interpret Wilson's various scores (see Tr. at 1241-42, 1340), Dr. James summarily rejected all of these opinions as mere "gut estimates" that "should be given little weight or importance" (James Rep. at 12 (internal quotation marks omitted)). The court finds her cursory analysis unpersuasive, particularly because she failed to interview any of Wilson's IQ examiners to determine if their conclusions were in fact based on their "gut" or were based on reasoned clinical judgments, which are indisputably essential to the interpretation of IQ tests.[33] (See Tr. at 1339-40.)

Finally, Dr. James was selective about her treatment of prorated IQ scores and raw data. She significantly discounted the score Wilson obtained on Dr. Popp's exam—his highest IQ score—because of evidence that Dr. Popp prorated this score (see id. 1219-21; supra n.30), and yet she did not place much weight on evidence that Wilson's score on Dr. Nagler's test—his lowest score—may have been prorated as well (see id. at 1324-25 (acknowledging that there was a dash in the box instead of a score for the "object assembly" subtest); id. at 1327 (agreeing that the absence of an object assembly score sheet was "pretty good evidence that Nagler prorated"). Dr. James testified that, because there was a page missing in Dr. Nagler's raw data, she did not have enough information to determine whether Dr. Nagler prorated the score. (Id. at 1355-56.) And that seems reasonable, but it is in serious tension with Dr. James's rather rigid view of raw data. (See Part III.A.4.) In her report, Dr. James argued that those "tests which lack raw data

[33] Dr. James suggested that the test examiners "weren't made available to [her]" (Tr. at 1339), but there is no indication in the record that this is true. Wilson's attorneys and his experts Dr. Olley and Dr. Woods spoke to a number of his IQ test examiners. (See, e.g., Olley Rep. at 8; Tr. at 1576-77, 1580.)

should be . . . given little weight in determining Mr. Wilson's intellectual functioning." (James Rep. at 10.) If indeed Dr. Nagler's report lacks sufficient raw data to clarify whether she prorated Wilson's IQ score—a point that Dr. James considered essential to the weight given Dr. Popp's test—then under Dr. James's own testimony, Dr. Nagler's score should be given "little weight." (Id.) Instead, Dr. James relied heavily upon Dr. Nagler's score and discounted Dr. Popp's. She cannot have it both ways.

5. Summary of Wilson's Intellectual Functioning

Even after taking into account the possibility of measurement error, the Flynn Effect, and (to a limited extent) the practice effect, Wilson's IQ scores are not indicative of significantly subaverage intellectual functioning. Seven of his eight Flynn-adjusted scores are at least 3 points above 70, the benchmark for mental retardation. The median and average of his scores are 77.19 and 76.44, respectively, both above what many courts have considered to be the cutoff. Seven of his eight scores have a 66% confidence interval beginning above 70, and the median and average of the bottom ends of the 66% confidence intervals are 74.21 and 73.59, respectively, suggesting that we can be at least 66% confident that his true IQ score lies higher than the benchmark of 70 (an average of 3.59 points higher). The only score that is suggestive of mental retardation—Wilson's score on Dr. Nagler's test—appears to be an outlier.

The clinical judgments of Wilson's test administrators support the court's analysis. None of these clinicians concluded that he suffered from mental retardation. And most of them believed that his observed scores represented an underestimate of his true intelligence. Dr. Nagler's notes on Wilson's behavior during her test, which suggest a poor attitude and a lack of effort on his part, further indicate that Wilson's score on her test was an unreliable outlier. In contrast, Dr. Denney's test (on which Wilson obtained a Flynn-adjusted 78.02 with a 66%

confidence interval of 75.90 to 80.14) is particularly reliable because: (1) only he administered the WAIS-IV, the "gold standard of IQ tests"; (2) the raw data for his test is indisputably complete and reveals no scoring errors; and (3) the test was administered eight and a half years after Wilson's previous test, diminishing (or eliminating) the potential for a practice effect.

With seven of his eight IQ scores and all of the opinions of his test administrators pointing away from mental retardation, Wilson is essentially left with the opinion of Dr. James, the one expert who performed a substantial analysis of his intellectual functioning and concluded that he satisfied prong one. The court finds her opinion inconsistent with the weight of the evidence, and has discussed a number of reasons to give it less weight than the opinions of Drs. Denney and Mapou, including her failure to administer her own IQ test and the inconsistent and selective nature of some of her findings.

For these reasons, and after a review of the full record in this case, the court concludes that Wilson has not satisfied his burden of proving that he more likely than not suffers from significantly subaverage intellectual functioning. He has thus failed to satisfy an indispensable prerequisite of the definition of mental retardation, see AAIDD 2010 Manual at 7, 27, 41; DSM-IV-TR at 49; see also Atkins, 536 U.S. at 318, and there is no need for the court to address the other requirements of the definition.

IV. CONCLUSION

The court holds that Wilson is not mentally retarded, and was not mentally retarded at the time of the crime. This does not mean that he will receive—or deserves to receive—the death penalty, but only that any such penalty would not violate the Federal Death Penalty Act or the Eighth Amendment. See 18 U.S.C. § 3596(c); Atkins, 536 U.S. at 318. The question of whether

Wilson is deserving of a death sentence shall be decided by a jury after a penalty phase trial.

SO ORDERED.

/s/ Nicholas G. Garaufis
NICHOLAS G. GARAUFIS
United States District Judge

Dated: Brooklyn, New York
February 7, 2013