No. _____
(Capital Case)

In the
Supreme Court of the United States

QUINTIN PHILLIPPE JONES, Petitioner,

v.

BOBBY LUMPKIN, Respondent.

On Petition for Writ of Certiorari to the Texas Court of Criminal Appeals

---

PETITION FOR WRIT OF CERTIORARI

APPENDIX

TABLE OF APPENDIX

IN THE COURT OF CRIMINAL APPEALS
OF TEXAS

NO. WR-57,299-02

EX PARTE QUINTIN PHILLIPPE JONES, Applicant

ON APPLICATION FOR POST-CONVICTION WRIT OF HABEAS CORPUS AND MOTION TO STAY THE EXECUTION FROM CAUSE NO. C-1-W011962-0744493-B IN CRIMINAL DISTRICT COURT NO. 1 TARRANT COUNTY

***Per curiam*.**

O R D E R

We have before us a subsequent application for a writ of habeas corpus filed pursuant to the provisions of Texas Code of Criminal Procedure Article 11.071 § 5, and a motion to stay Applicant's execution.[1]

In February 2001, a jury convicted Applicant of the September 1999 intentional

[1] All references to "Articles" in this order refer to the Texas Code of Criminal Procedure unless otherwise specified.

killing of his elderly great aunt committed in the course of robbing or attempting to rob her. *See* TEX. PENAL CODE § 19.03(a). Based on the jury's answers to the special issues submitted pursuant to Article 37.071, the trial court sentenced Applicant to death. Art. 37.071 § 2(g). This Court affirmed Applicant's conviction and sentence on direct appeal. *Jones v. State*, 119 S.W.3d 766 (Tex. Crim. App. 2003). We also denied relief on Applicant's initial writ of habeas corpus application. *Ex parte Jones*, No. WR-57,299-01 (Tex. Crim. App. Sept. 14, 2005) (not designated for publication).

On May 6, 2021, Applicant filed in the trial court the instant writ application in which he raises three claims. In his first two claims, Applicant asserts that his death sentence was obtained in violation of the Fourteenth Amendment's due process clause because it was based on false and misleading scientific evidence. In his third claim, Applicant asserts that he may be intellectually disabled and, therefore, cannot constitutionally be executed.

We have reviewed the application and find that Applicant has failed to make a *prima facie* showing on any of his allegations. Therefore, the allegations do not satisfy the requirements of Article 11.071 § 5. Accordingly, we dismiss the application as an abuse of the writ without reviewing the merits of the claim raised. Art. 11.071 § 5(c). We deny Applicant's motion to stay his execution.

IT IS SO ORDERED THIS THE 12th DAY OF MAY, 2021.

Do Not Publish

IC
500,000

& MURDER & AGG ROBB-SBI
& AGG ROBB-BI-ELDERLY

NAME QUINTIN PHILLIPPE JONES
aka: QUINTON JONES

OFFENSE CAPITAL MUR

ADDRESS 3521 BAYLOR

DATE 09-11-99

FT WORTH TX 00000

I. P. BERTHENA BRYANT

RACE B SEX M AGE 20 DOB 07-15-79

C. C.

CASE NO. 0744493 FILED: (DATE) 09-17-99

AGENCY FORT WORTH PD

PC HAS BEEN DETERMINED
TRANSFER: 04/30/98 COURT DATE

OFFENSE NO. 99606040 COURT CDC1

INDICTMENT NO. 0744493 D

IN THE NAME AND BY AUTHORITY OF THE STATE OF TEXAS:

THE GRAND JURORS OF TARRANT COUNTY, TEXAS, duly elected, tried, empaneled, sworn and charged to inquire of offenses committed in Tarrant County, in the State of Texas, upon their oaths do present in and to the * * * * * * * * * * CRIMINAL DISTRICT COURT NO. 4 of said County that * * QUINTIN PHILLIPPE JONES aka: QUINTON JONES hereinafter called Defendant, in the County of Tarrant and State aforesaid, on or about the 11TH day of SEPTEMBER 1999, did

NG

THEN AND THERE INTENTIONALLY CAUSE THE DEATH OF AN INDIVIDUAL, BERTHENA BRYANT, BY STRIKING HER WITH A DEADLY WEAPON, TO-WIT: A BASEBALL BAT, THAT IN THE MANNER OF ITS USE OR INTENDED USE WAS CAPABLE OF CAUSING DEATH OR SERIOUS BODILY INJURY, AND THE SAID DEFENDANT WAS THEN AND THERE IN THE COURSE OF COMMITTING OR ATTEMPTING TO COMMIT ROBBERY OF BERTHENA BRYANT,

with

COUNT TWO: AND, IT IS FURTHER PRESENTED IN AND TO SAID COURT, THAT THE SAID DEFENDANT, IN THE COUNTY OF TARRANT AND STATE AFORESAID, ON OR ABOUT THE 11TH DAY OF SEPTEMBER, 1999, DID THEN AND THERE INTENTIONALLY OR KNOWINGLY CAUSE THE DEATH OF AN INDIVIDUAL, BERTHENA BRYANT, BY STRIKING HER WITH A DEADLY WEAPON, TO-WIT: A BASEBALL BAT, THAT IN THE MANNER OF ITS USE OR INTENDED USE WAS CAPABLE OF CAUSING DEATH OR SERIOUS BODILY INJURY,

PARAGRAPH TWO: AND IT IS FURTHER PRESENTED IN AND TO SAID COURT THAT THE SAID DEFENDANT IN THE COUNTY OF TARRANT AND STATE AFORESAID ON OR ABOUT THE 11TH DAY OF SEPTEMBER, 1999, DID THEN AND THERE INTENTIONALLY WITH THE INTENT TO CAUSE SERIOUS BODILY INJURY TO BERTHENA BRYANT, COMMIT AN ACT CLEARLY DANGEROUS TO HUMAN LIFE, NAMELY, STRIKING HER WITH A DEADLY WEAPON, TO-WIT: A BASEBALL BAT, THAT IN THE MANNER OF ITS USE OR INTENDED USE WAS CAPABLE OF CAUSING DEATH OR SERIOUS BODILY INJURY, WHICH CAUSED THE DEATH OF BERTHENA BRYANT,

COUNT THREE: AND, IT IS FURTHER PRESENTED IN AND TO SAID COURT, THAT THE SAID DEFENDANT, IN THE COUNTY OF TARRANT AND STATE AFORESAID, ON OR ABOUT THE 11TH DAY OF SEPTEMBER, 1999, DID THEN AND THERE INTENTIONALLY OR KNOWINGLY, WHILE IN THE COURSE OF COMMITTING THEFT OF PROPERTY AND WITH INTENT TO OBTAIN OR MAINTAIN CONTROL OF SAID PROPERTY, CAUSE SERIOUS BODILY INJURY TO BERTHENA BRYANT, BY STRIKING HER WITH A DEADLY WEAPON, TO-WIT: A BASEBALL BAT, THAT IN THE MANNER OF ITS USE OR INTENDED USE WAS CAPABLE OF CAUSING DEATH OR SERIOUS BODILY INJURY,

COUNT FOUR: AND, IT IS FURTHER PRESENTED IN AND TO SAID COURT, THAT THE SAID

(CONTINUED)

DEFENDANT, IN THE COUNTY OF TARRANT AND STATE AFORESAID, ON OR ABOUT THE 11TH DAY OF SEPTEMBER, 1999, DID THEN AND THERE INTENTIONALLY OR KNOWINGLY, WHILE IN THE COURSE OF COMMITTING THEFT OF PROPERTY AND WITH INTENT TO OBTAIN OR MAINTAIN CONTROL OF SAID PROPERTY, CAUSE BODILY INJURY TO BERTHENA BRYANT A PERSON 65 YEARS OF AGE OR OLDER, BY STRIKING HER WITH A BASEBALL BAT, THAT IN THE MANNER OF ITS USE OR INTENDED USE WAS CAPABLE OF CAUSING DEATH OR SERIOUS BODILY INJURY,

Filed (Clerk's use only)

FILED
THOMAS A. WILDER, DIST. CLERK
TARRANT COUNTY, TEXAS

DEC 01 1999

Time 10:00 a

By _____ Deputy

AGAINST THE PEACE AND DIGNITY OF THE STATE.

Criminal District Attorney

Foreman of the Grand Jury

CASE NO. 0744493D

THE STATE OF TEXAS	§	IN THE CRIMINAL
VS.	§	DISTRICT COURT ONE OF
QUINTIN PHILLIPPE JONES aka: QUINTON JONES	§	TARRANT COUNTY, TEXAS

CAPITAL JUDGMENT

On FEBRUARY 14, 2001, this cause was called for trial and the State by her Criminal District Attorney, Assistants CHRISTY JACK, and BEN LEONARD and the attorneys for the Defendant, QUINTIN PHILLIPPE JONES aka: QUINTON JONES, Honorables LARRY MOORE and J. REX BARNETT, announced ready for trial; and the State having made known that it would seek the Death Penalty in this cause and the Defendant having been heretofore arraigned; and, it appearing to the Court that the Defendant was mentally competent and the Defendant having been charged in the indictment with Capital Murder; thereupon, a Jury of good and lawful men and women, to-wit: Leslie D. Skinner, Foreperson, and eleven others, was duly selected, impaneled and sworn as the law directs, and the said Criminal District Attorney read to the Jury, Count One of the indictment herein, and the Defendant entered his plea of Not Guilty to Count One, of the indictment, hereto; and the Jury, after hearing the evidence, and being duly charged by the Court, retired to consider its verdict, and after deliberation, returned into open Court on the 21st day of FEBRUARY, 2001, the following verdict, to-wit:

VERDICT FORM

We the Jury, find the Defendant, QUINTIN PHILLIPPE JONES aka: QUINTON JONES, guilty of the offense of Capital Murder, as charged in the indictment.

Signed: *Leslie D. Skinner*
Foreperson of the Jury

The parties announced ready for the second phase of the trial, and the Jury, having heard all the evidence, and being duly charged by the Court, retired to consider its verdict, and after due deliberation, returned into open court, on the 26TH day of FEBRUARY, 2001, their answers to the following Special Issues, and their verdict:

ISSUE NUMBER 1

Do you find from the evidence beyond a reasonable doubt that there is a probability that the Defendant would commit criminal acts of violence that would constitute a continuing threat to society?

In your verdict you will answer "Yes" or "No" Answer: YES

ISSUE NUMBER 2

Taking into consideration all of the evidence, including the circumstances of the offense, the Defendant's character and background, and the personal moral culpability of the Defendant, do you find that there is a sufficient mitigating circumstance or circumstances to warrant that a sentence of life imprisonment rather than a death sentence be imposed?

In you verdict you will answer "Yes" or "No" Answer: NO

VERDICT FORM

We, the Jury, having unanimously agreed upon the answer to the foregoing issues do hereby return the same into court as our verdict.

Signed: *Leslie D. Skinner*
Foreperson of the Jury

After an individual poll of the Jurors, the Court duly accepted the verdicts and ORDERED the same to be filed.

The Jury having answered Issue Number One "YES" and Issue Number Two, "NO", it being mandatory that the punishment be death, the Court assessed the punishment at Death.

The Defendant, QUINTIN PHILLIPPE JONES aka: QUINTON JONES, was asked by the Court, whether he had anything to say why sentence should not be pronounced against him, and the Defendant answered nothing in bar thereof;

The Court proceeded, in the presence of the said Defendant JINTIN PHILLIPPE JONES aka: QUINTON JONES, and his counsel of record, to pronounce sentence against him as follows:

Quintin Jones, the jury having found you guilty of the capital murder of Berthena Bryant, and having returned a unanimous verdict to Issue Numbers 1 and 2, the Court sentences you to death by lethal injection.

It is, therefore, the order, judgment, and decree of this Court that you be remanded to the custody of the Sheriff of this county to be delivered to the Institutional Division of the Texas Department of Criminal Justice, where you shall be continuously confined until 6:00 p.m. on a date to be determined when the proper authorities shall administer lethal injections sufficient to cause your death.

HON. SHAREN WILSON
PRESIDING JUDGE
CRIMINAL DISTRICT COURT ONE
TARRANT COUNTY, TEXAS

MARCH 5, 2001
Date Signed

NO. 0744493D

THE STATE OF TEXAS	§	IN THE CRIMINAL DISTRICT
	§	
VS.	§	COURT NUMBER ONE
	§	
QUINTON JONES	§	TARRANT COUNTY, TEXAS

COURT'S CHARGE

MEMBERS OF THE JURY:

The Defendant, Quinton Jones, stands charged by indictment with the offense of capital murder, alleged to have been committed on or about the 11th day of September, 1999, in Tarrant County, Texas. To this charge, the Defendant has pleaded not guilty.

Our law provides that a person commits murder when he intentionally or knowingly causes the death of an individual.

A person commits capital murder when such person intentionally commits the murder in the course of committing or attempting to commit the offense of robbery.

A person commits robbery if in the course of committing theft and with intent to obtain or maintain control of the property of another, he intentionally or knowingly causes bodily injury to another.

In the course of committing theft means conduct that occurs in an attempt to commit, during the commission, or in immediate flight after the attempt or commission of theft.

A person commits theft if he unlawfully appropriates property with intent to deprive the owner of property. Appropriation of property is unlawful if it is without the owner's effective consent.

Appropriation, or appropriate, means to acquire or otherwise exercise control over property other than real property.

Property means tangible or intangible personal property including anything severed from land; or a document, including money, that represents or embodies anything of value.

Deprive means to withhold property from the owner permanently or for so extended a period of time that a major portion of the

use or enjoyment of the property is lost to the owner.

Consent means assent in fact, whether express or apparent.

Effective consent includes consent by a person legally authorized to act for the owner. Consent is not effective if induced by force, threat or fraud.

Coercion means a threat, however communicated, to commit an offense.

Owner means a person who has title to the property, possession of the property, whether lawful or not, or a greater right to possession of the property than the actor.

Bodily injury means physical pain, illness, or any impairment of physical condition.

Serious bodily injury means bodily injury that creates a substantial risk of death or that causes death, serious permanent disfigurement, or protracted loss or impairment of the function of any bodily member or organ.

A deadly weapon means anything manifestly designed, made, or adapted for the purpose of inflicting death or serious bodily injury.

A person acts intentionally, or with intent, with respect to the result of his conduct when it is his conscious objective or desire to cause the result.

With respect to the offense of murder, a person acts knowingly, or with knowledge, with respect to the nature of his conduct or to a result of his conduct when he is aware of the nature of his conduct or that the circumstances exist. A person acts knowingly, or with knowledge, with respect to the result of his conduct when he is aware that his conduct is reasonably certain to cause the result.

You are further charged as the law in this case that the State is not required to prove the exact date alleged in the indictment, but may prove the offense, if any, to have been committed at any time prior to the presentment of the indictment.

Now, if you find from the evidence beyond a reasonable doubt that on or about the 11th day of September, 1999, in Tarrant County, Texas, the Defendant, Quinton Jones, did then and there intentionally cause the death of an individual, Berthena Bryant, by striking her with a deadly weapon, to-wit: a baseball bat, that in the manner of its use or intended use was capable of causing death or serious bodily injury, and the said defendant was then and there in the course of committing or attempting to commit robbery of Berthena Bryant, then you will find the Defendant guilty of capital murder as charged in the indictment.

Unless you so find beyond a reasonable doubt, or if you have a reasonable doubt thereof, you will find the defendant not guilty of capital murder, as charged in the indictment, and you will consider whether the defendant is guilty of the offense of murder.

Now, if you find from the evidence beyond a reasonable doubt that on or about the 11th day of September, 1999, in Tarrant County, Texas, the Defendant, Quinton Jones, did then and there intentionally or knowingly cause the death of an individual, Berthena Bryant, by striking her with a deadly weapon, to-wit: a baseball bat, that in the manner of its use or intended use was capable of causing death or serious bodily injury, then you will find the Defendant guilty of murder.

Unless you so find beyond a reasonable doubt, or if you have a reasonable doubt thereof, you will find the defendant not guilty.

If you should find from the evidence beyond a reasonable doubt that the defendant is either guilty of capital murder or murder, but you have a reasonable doubt as to which offense he is guilty, then you should resolve that doubt in the defendant's favor, and in such event you will find the defendant guilty of the lesser offense of murder.

Voluntary intoxication does not constitute a defense to the commission of a crime. Intoxication means substantial impairment of mental or physical capacity resulting from introduction of any

substance into the body.

You are instructed that if there is any testimony before you in this case regarding the defendant's having committed offenses other than the offense alleged against him in the indictment in this case, you cannot consider said testimony for any purpose unless you find and believe beyond a reasonable doubt that the defendant committed such other offenses, if any were committed, and even then you may only consider the same in determining the intent of the Defendant, if any, in connection with the offense alleged against him in the indictment in this case, and for no other purpose.

You are instructed that our law provides that a person is entitled to the assistance of counsel prior to and during any questioning which takes place while the person is in the custody of a peace officer if he makes an unambiguous request for a lawyer. Before questioning can continue, the person must reinitiate the interview by indicating a willingness to continue the interview. Therefore, unless you believe from the evidence beyond a reasonable doubt that the alleged confession or statement introduced into evidence resulted from a free and voluntary waiver of his right to counsel without compulsion or persuasion, or if you have a reasonable doubt thereof, you shall not consider such alleged statement or confession for any purpose nor any evidence obtained as a result thereof.

You are instructed that our law provides that a defendant may testify in his own behalf if he chooses to do so. This, however, is a right accorded to a defendant, and in the event he chooses not to testify, that fact cannot be taken as a circumstance against him. In this case, the Defendant has chosen not to testify and you are instructed that you cannot and must not refer or allude to that fact throughout your deliberations or take it into consideration for any purpose whatsoever as a circumstance against him.

You have been permitted to take notes during the testimony in

his case. In the event any of you took notes, you may rely on your notes during your deliberations. However, you may not share your notes with the other jurors and you should not permit the other jurors to share their notes with you. You may, however, discuss the contents of your notes with the other jurors. You shall not use your notes as authority to persuade your fellow jurors. In your deliberations, give no more and no less weight to the views of a fellow juror just because that juror did or did not take notes. Your notes are not official transcripts. They are personal memory aids, just like the notes of the judge and the notes of the lawyers. Notes are valuable as a stimulant to your memory. On the other hand, you might make an error in observing or you might make a mistake in recording what you have seen or heard. Therefore, you are not to use your notes as authority to persuade fellow jurors of what the evidence was during the trial.

Occasionally, during jury deliberations, a dispute arises as to the testimony presented. If this should occur in this case, you shall inform the court and request that the court read the portion of disputed testimony to you from the official transcript. You shall not rely on your notes to resolve the dispute because those notes, if any, are not official transcripts. The dispute must be settled by the official transcript, for it is the official transcript, rather than any juror's notes, upon which you must base your determination of the facts and, ultimately, your verdict in this case.

Your verdict must be by a unanimous vote of all members of the jury. In deliberating on this case, you shall consider the charge as a whole and you must not refer to nor discuss any matters not in evidence.

In all criminal cases, the burden of proof is on the State. The burden of proof rests upon the State throughout the trial and never shifts to the Defendant.

The indictment in this case is no evidence whatsoever of the

guilt of the Defendant. It is a mere pleading that is necessary in order to bring this case into court for trial and you will not consider it for any purpose.

All persons are presumed to be innocent and no person may be convicted of an offense unless each element of the offense is proved beyond a reasonable doubt. The fact that a defendant has been arrested, confined, indicted for, or otherwise charged with an offense gives no rise to any inference of guilt at his trial.

You are the exclusive judges of the facts proved, of the credibility of the witnesses, and of the weight to be given their testimony. But you are bound to receive the law from the Court, which is given in these written instructions, and be governed thereby.

After you retire to the jury room, you should select one of your members as your foreman. It is his or her duty to preside at your deliberations, to vote with you, and when you have reached a unanimous verdict, to certify to your verdict by using the attached form and signing the same as the foreman.

You may communicate with this Court in writing through the bailiff who has you in charge. Your written communication must be signed by the foreman. Do not attempt to talk to the bailiff, the attorneys, or the Court regarding any question you may have concerning the trial of the case. After you have reached a unanimous verdict or if you desire to communicate with the Court, please use the jury call button on the wall and one of the bailiffs will respond.

SHAREN WILSON, JUDGE
Criminal District Court No. 1
Tarrant County, Texas

VERDICT FORM

We, the jury, find the Defendant, Quinton Jones, guilty of the offense of capital murder, as charged in the indictment.

FOREMAN

-OR-

We, the jury, find the Defendant, Quinton Jones, guilty of the offense of murder.

FOREMAN

-OR-

We, the jury, find the Defendant, Quinton Jones, not guilty.

FOREMAN

NO. 0744493D

THE STATE OF TEXAS	§	IN THE CRIMINAL DISTRICT
	§	
VS.	§	COURT NUMBER ONE
	§	
QUINTON JONES	§	TARRANT COUNTY, TEXAS

COURT'S CHARGE

MEMBERS OF THE JURY:

You have found the defendant, Quinton Jones, guilty of the offense of capital murder. You are instructed that the defendant shall be punished by imprisonment in the institutional division of the Texas Department of Criminal Justice for life or by death. In order for the Court to assess the proper punishment, the Court submits to you issues.

ISSUE NUMBER 1: Whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society.

You shall return a special verdict of "yes" or "no" to Issue Number 1. The prosecution has the burden of proving that the answer to Issue Number 1 should be "yes", and it must do so by proving Issue Number 1 beyond a reasonable doubt, and if it fails to do so, you must answer Issue Number 1 "no".

In deliberating on Issue Number 1, you shall consider all evidence admitted at the guilt or innocence stage and the punishment stage, including evidence of the defendant's background or character or the circumstances of the offense that militates for or mitigates against the imposition of the death penalty.

You may not answer Issue Number 1 "yes" unless you agree unanimously. You may not answer Issue Number 1 "no" unless 10 or more jurors agree. The members of the jury need not agree on what particular evidence supports a negative answer to Issue Number 1.

If the jury answers Issue Number 1 "yes", then you shall answer Issue Number 2; otherwise, do not answer Issue Number 2.

ISSUE NUMBER 2: Whether, taking into consideration all of the evidence, including the circumstances of the offense, the defendant's character and background, and the personal moral culpability of the defendant, there is a sufficient mitigating circumstance or circumstances to warrant that a sentence of life imprisonment rather than a death sentence be imposed.

You shall return a special verdict of "yes" or "no" to Issue Number 2. You are instructed that you may not answer Issue Number 2 "no" unless you agree unanimously. You may not answer Issue Number 2 "yes" unless ten or more jurors agree. The members of the jury need not agree on what particular evidence supports an affirmative finding on the issue. In deliberating on Issue Number 2, you shall consider mitigating evidence to be evidence that a juror might regard as reducing the defendant's moral blameworthiness.

In arriving at the answers to the above issues, it will not be proper for you to fix the same by lot, chance, or any other method than a full, fair, and free exercise of the opinion of the individual jurors.

Under the law applicable in this case, if the defendant is sentenced to imprisonment in the institutional division of the Texas Department of Criminal Justice for life, the defendant will become eligible for release on parole, but not until the actual time served by the defendant equals 40 years, without consideration of good conduct time. It cannot accurately be predicted how the parole laws might be applied to this defendant if the defendant is sentenced to a term of imprisonment for life because the application of those laws will depend on decisions made by prison and parole authorities, but eligibility for parole does not guarantee that parole will be granted.

You are instructed that our law provides that a defendant may testify in his own behalf if he chooses to do so. This, however, is a privilege accorded to a defendant, and in the event he chooses

not to testify, that fact cannot be taken as a circumstance against him. In this case, the defendant has chosen not to testify and you are instructed that you cannot and must not refer or allude to that fact throughout your deliberations or take it into consideration for any purpose whatsoever as a circumstance against him.

You are instructed that our law provides that before you may consider a statement of the defendant as evidence against him, you must find that the defendant knowingly, intelligently and voluntarily waived the right to counsel. Therefore, unless you find beyond a reasonable doubt that prior to giving the statement offered into evidence, and during the course of giving such statement, the defendant knowingly, intelligently, and voluntarily waived the right to have assistance of counsel, or if you have a reasonable doubt as to whether the defendant waived this right, you shall not consider such statement or any evidence obtained as a result of such statement, for any purpose whatsoever.

You are instructed that if there is any testimony before you in this case regarding the defendant's having committed offenses or bad acts other than the offense for which you have found him guilty, you cannot consider said testimony for any purpose unless you find and believe beyond a reasonable doubt that the defendant committed such other offenses or bad acts, if any were committed, and even then you may only consider the same in determining the punishment which you will assess against the defendant in this case.

You have been permitted to take notes during the testimony in this case. In the event any of you took notes, you may rely on your notes during your deliberations. However, you may not share your notes with the other jurors and you should not permit the other jurors to share their notes with you. You may, however, discuss the contents of your notes with the other jurors. You shall not use your notes as authority to persuade your fellow jurors. In your deliberations, give no more and no less weight to

the views of a fellow juror just because that juror did or did not take notes. Your notes are not official transcripts. They are personal memory aids, just like the notes of the judge and the notes of the lawyers. Notes are valuable as a stimulant to your memory. On the other hand, you might make an error in observing or you might make a mistake in recording what you have seen or heard. Therefore, you are not to use your notes as authority to persuade fellow jurors of what the evidence was during the trial.

Occasionally, during jury deliberations, a dispute arises as to the testimony presented. If this should occur in this case, you shall inform the court and request that the court read the portion of disputed testimony to you from the official transcript. You shall not rely on your notes to resolve the dispute because those notes, if any, are not official transcripts. The dispute must be settled by the official transcript, for it is the official transcript, rather than any juror's notes, upon which you must base your determination of the facts and, ultimately, your verdict in this case.

After argument of counsel, you will retire to the jury room to deliberate. Any further communication must be in writing signed by your foreman through the bailiff to the Court. When you have reached a verdict, you may use the attached forms to indicate your answers to the Issues. Your foreman should sign the appropriate form certifying to your verdict.

JUDGE SHAREN WILSON
Criminal District Court No. 1
Tarrant County, Texas

Now, bearing in mind the foregoing instructions, you will answer the following issues:

ISSUE NUMBER 1

Do you find from the evidence beyond a reasonable doubt that there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society?

In your verdict, you will answer "yes" or "no".

ANSWER: Yes

If your answer to Issue Number 1 is "yes", then you will answer Issue Number 2; otherwise, you will not answer Issue Number 2.

ISSUE NUMBER 2

Taking into consideration all of the evidence, including the circumstances of the offense, the defendant's character and background, and the personal moral culpability of the defendant, do you find that there is a sufficient mitigating circumstance or circumstances to warrant that a sentence of life imprisonment rather than a death sentence be imposed?

In your verdict, you will answer "yes" or "no".

ANSWER: No

We, the jury, having agreed upon the answers to the foregoing issues, do hereby return the same into court as our verdict.

FOREMAN

CAUSE NO. 0744493D

THE STATE OF TEXAS	§	**IN THE CRIMINAL DISTRICT**
	§	
vs.	§	**COURT NO. 1 OF**
	§	
QUINTIN PHILLIPPE JONES	§	**TARRANT COUNTY, TEXAS**

ORDER SETTING EXECUTION DATE

The Court has reviewed the State's Motion for Court to Enter Order Setting Execution Date filed on October 12, 2020, and finds that the motion should be **GRANTED** and a date of execution be set in this case.

On February 21, 2001, a Tarrant County jury convicted Defendant for the capital murder of his eighty-three-year-old great aunt, Berthena Bryant, whom he beat to death with a baseball bat and robbed. CR 3: 397. On February 26, 2001, the jury returned an affirmative answer to the future-dangerousness special issue and a negative answer to the mitigation special issue, and this Court sentenced Defendant to death by lethal injection. CR 3: 408.

The Court of Criminal Appeals of Texas affirmed Defendant's conviction and sentence on direct appeal on November 5, 2003, and the Supreme Court of the United States denied his petition for a writ of certiorari on June 14, 2004. *Jones v. State*, 119 S.W.3d 766 (Tex. Crim. App. 2003), *cert. denied*, 542 U.S. 905 (2004).

On September 14, 2005, the Court of Criminal Appeals denied Defendant's state

application for a writ of habeas corpus. *Ex parte Jones*, 2005 WL 2220030 (Tex. Crim. App. Sept. 14, 2005) (unpublished per curiam order); and subsequently, on September 14, 2006, Defendant untimely filed his federal petition for writ of habeas corpus, and the United States District Court for the Northern District of Texas, Fort Worth Division, dismissed it as time-barred. *Jones v. Quarterman*, 2007 WL 2756755 (N.D. Tex. Sept. 21, 2007) (unpublished mem. op. & order). The district court later appointed Defendant new counsel and vacated its dismissal to give Defendant an opportunity to respond. *Jones v. Quarterman*, 2008 WL 4166850 (N.D. Tex. Sept. 10, 2008) (unpublished order). After Defendant responded, the district court again dismissed his petition as time-barred. *Jones v. Quarterman*, 2009 WL 559959 (N.D. Tex. Mar. 4, 2009) (unpublished mem. op. & order). Defendant appealed, and the United States Court of Appeals for the Fifth Circuit vacated and remanded for reconsideration in light of the principle of equitable tolling. *Jones v. Thaler*, 2010 WL 2464998 (5th Cir. June 17, 2010) (per curiam) (unpublished). Ultimately, the district court reversed course, and Defendant filed an amended petition. *Jones v. Stephens*, 998 F. Supp. 2d 529 (N.D. Tex. 2014). On January 13, 2016, the district court denied Defendant's claims for relief and denied a certificate of appealability on all of his claims. *Jones v. Stephens*, 157 F. Supp. 3d 623 (N.D. Tex. 2016).

The Fifth Circuit granted Defendant a certificate of appealability on his claim that his Fifth Amendment rights were violated by admission of an unmirandized

confession at the punishment phase of his trial. *Jones v. Davis*, 927 F.3d 365 (5th Cir. 2019). On June 18, 2019, the Fifth Circuit affirmed the district court's judgment. *Id.* Lastly, the Supreme Court of the United States denied Defendant's petition for a writ of certiorari on March 23, 2020. *Jones v. Davis*, 140 S. Ct. 2519 (2020).

IT IS THEREFORE EVIDENT that Defendant has exhausted his avenues for relief through the state and federal courts, and further there are no stays of execution in effect in this case.

ACCORDINGLY, IT IS HEREBY ORDERED that the Defendant, Quintin Phillippe Jones, who has been adjudged to be guilty of capital murder as charged in the indictment and whose punishment has been assessed by the verdict of the jury and judgment of the Court at **DEATH**, shall be kept or taken into the custody of the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice until the **19th day of May 2021**, upon which day, at the Correctional Institutions Division of the Texas Department of Criminal Justice, at some time after the hour of six o'clock p.m., in a room designated by the Correctional Institutions Division of the Texas Department of Criminal Justice and arranged for the purpose of execution, the said Director, acting by and through the executioner designated by said Director, as provided by law, is hereby commanded, ordered and directed to carry out this sentence of death by intravenous injection of a substance or substances in a lethal quantity sufficient to cause the death of the Defendant, Quintin Phillippe Jones, until

Quintin Phillippe Jones is dead. Such procedure shall be determined and supervised by the said Director of the Correctional Institutions Division of the Texas Department of Criminal Justice.

IT IS FURTHER ORDERED that the **Clerk of this Court** shall issue and deliver to the **Sheriff of Tarrant County, Texas**, a **Death Warrant** in accordance with this sentence and Order, directed to the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice, at Huntsville, Texas, commanding the said Director, to put into execution the Judgment of Death against the said Quintin Phillippe Jones.

The Sheriff of Tarrant County, Texas IS HEREBY ORDERED, upon receipt of said Death Warrant, to deliver said Warrant to the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice, Huntsville, Texas together with Defendant, Quintin Phillippe Jones.

IT IS FURTHER ORDERED that the **Clerk of this Court** shall deliver a copy of this order, by first-class mail, e-mail, or fax not later than the second business day after the Court enters the order, to:

1. The attorney who represented the Defendant in the most recently concluded stage of a state or federal post-conviction proceeding, Michael Mowla, michael@mowlalaw.com. *See* TEX. CODE CRIM. PROC. art. 43.141(b-1).
2. Benjamin Wolff, Director, Office of Capital and Forensic Writs, Benjamin.Wolff@ocfw.texas.gov. *See Id.*
3. Rachel Patton, Assistant Attorney General,

Rachel.Patton@oag.texas.gov.

4. The Post-Conviction Section of the Tarrant County Criminal District Attorney's Office, ccaappellatealerts@tarrantcountytx.gov.

SIGNED AND ENTERED this 18th day of November 2020.

HON. ELIZABETH BEACH
JUDGE PRESIDING
CRIMINAL DISTRICT COURT NO. 1

0744493D

Death Warrant and Execution Order for QUINTIN PHILLIPPE JONES was hand-delivered by the Sheriff of Tarrant County to Texas Department of Criminal Justice, Classification and Records on this 19th day of November, 2020.

Received by:

Debra Burleson for
Bryan Collier, Executive Director
Texas Department of Criminal Justice

Delivered by:

Sheriff

FILED
THOMAS A WILDER, DIST. CLERK
TARRANT COUNTY, TEXAS
NOV 23 2020
TIME 9:36 am
DEPUTY

THE STATE OF TEXAS	§	IN THE CRIMINAL DISTRICT
	§	
VS.	§	COURT ONE OF
	§	
CAUSE NO. 0744493D	§	TARRANT COUNTY, TEXAS
	§	
QUINTIN PHILLIPPE JONES	§	

DEATH WARRANT

To the Director of the Correctional Institutions Division of the Texas Department Of Criminal Justice at Huntsville, Texas, or in case of his death, disability or absence, the Warden of the Huntsville Unit of the Correctional Institutions Division of the Texas Department of Criminal Justice or in the event of the death or disability or absence of both the Director of the Correctional Institutions Division of the Texas Department Of Criminal Justice and the Warden of the Correctional Institutions Division of the Texas Department Of Criminal Justice, to such person appointed by the Board of Directors of the Correctional Institutions Division of the Texas Department Of Criminal Justice, Greetings:

Whereas, on the 21ST day of FEBRUARY, A.D. 2001, in the CRIMINAL District Court ONE of Tarrant County, Texas, QUINTIN PHILLIPPE JONES was duly and legally convicted of the crime of Capital Murder, as fully appears in the judgment of said Court entered upon the minutes of said court as follows, to-wit: Judgment attached and,

Whereas, on the 26TH day of FEBRUARY, A.D., 2001 the said Court pronounced sentence upon the said QUINTIN PHILLIPPE JONES in accordance with said judgment fixing the time for the execution of the said QUINTIN PHILLIPPE JONES for any time after the hour of 6:00 p.m. on WEDNESDAY, the 19TH day of MAY, A.D., 2021, as fully appears in the sentence of the Court and entered upon the minutes of said Court as follows, to-wit: Sentence attached.

These are therefore to command you to execute the aforesaid judgment and sentence any time after the hour of 6:00 p.m. on WEDNESDAY, the 19TH day of MAY, A.D., 2021, by intravenous injection of substance or substances in a lethal quantity sufficient to cause death and until the said QUINTIN PHILLIPPE JONES is dead.

Herein fail not, and due return make hereof in accordance with law.

Witness my signature and seal of office on this the 18TH day of NOVEMBER, A.D., 2020.

Issued under my hand and seal of Office in the City of Fort Worth, Tarrant County Texas this 18TH day of NOVEMBER, 2020.

DISTRICT COURT OF TARRANT COUNTY, TEXAS

THOMAS A. WILDER,
CLERK OF THE DISTRICT COURTS OF
TARRANT COUNTY, TEXAS

BY _____, Deputy

RETURN OF THE DIRECTOR OF THE TEXAS DEPARTMENT OF CORRECTIONS

Came to hand, this the _____ day of _____, _____ and executed the ____ day of _____, ____ by the death of

QUINTIN PHILLIPPE JONES

DISPOSITION OF BODY:

DATE:

TIME:

DIRECTOR OF TEXAS DEPARTMENT OF CORRECTIONS

BY: _____

CAUSE NO. 0744493D

THE STATE OF TEXAS	§	**IN THE CRIMINAL DISTRICT**
	§	
vs.	§	**COURT NO. 1 OF**
	§	
QUINTIN PHILLIPPE JONES	§	**TARRANT COUNTY, TEXAS**

DUPLICATE ORDER SETTING EXECUTION DATE

The Court has reviewed the State's Motion for Court to Enter Order Setting Execution Date filed on October 12, 2020, and finds that the motion should be **GRANTED** and a date of execution be set in this case.

On February 21, 2001, a Tarrant County jury convicted Defendant for the capital murder of his eighty-three-year-old great aunt, Berthena Bryant, whom he beat to death with a baseball bat and robbed. CR 3: 397. On February 26, 2001, the jury returned an affirmative answer to the future-dangerousness special issue and a negative answer to the mitigation special issue, and this Court sentenced Defendant to death by lethal injection. CR 3: 408.

The Court of Criminal Appeals of Texas affirmed Defendant's conviction and sentence on direct appeal on November 5, 2003, and the Supreme Court of the United States denied his petition for a writ of certiorari on June 14, 2004. *Jones v. State*, 119 S.W.3d 766 (Tex. Crim. App. 2003), *cert. denied*, 542 U.S. 905 (2004).

On September 14, 2005, the Court of Criminal Appeals denied Defendant's state

application for a writ of habeas corpus. *Ex parte Jones*, 2005 WL 2220030 (Tex. Crim. App. Sept. 14, 2005) (unpublished per curiam order); and subsequently, on September 14, 2006, Defendant untimely filed his federal petition for writ of habeas corpus, and the United States District Court for the Northern District of Texas, Fort Worth Division, dismissed it as time-barred. *Jones v. Quarterman*, 2007 WL 2756755 (N.D. Tex. Sept. 21, 2007) (unpublished mem. op. & order). The district court later appointed Defendant new counsel and vacated its dismissal to give Defendant an opportunity to respond. *Jones v. Quarterman*, 2008 WL 4166850 (N.D. Tex. Sept. 10, 2008) (unpublished order). After Defendant responded, the district court again dismissed his petition as time-barred. *Jones v. Quarterman*, 2009 WL 559959 (N.D. Tex. Mar. 4, 2009) (unpublished mem. op. & order). Defendant appealed, and the United States Court of Appeals for the Fifth Circuit vacated and remanded for reconsideration in light of the principle of equitable tolling. *Jones v. Thaler*, 2010 WL 2464998 (5th Cir. June 17, 2010) (per curiam) (unpublished). Ultimately, the district court reversed course, and Defendant filed an amended petition. *Jones v. Stephens*, 998 F. Supp. 2d 529 (N.D. Tex. 2014). On January 13, 2016, the district court denied Defendant's claims for relief and denied a certificate of appealability on all of his claims. *Jones v. Stephens*, 157 F. Supp. 3d 623 (N.D. Tex. 2016).

The Fifth Circuit granted Defendant a certificate of appealability on his claim that his Fifth Amendment rights were violated by admission of an unmirandized

confession at the punishment phase of his trial. *Jones v. Davis*, 927 F.3d 365 (5th Cir. 2019). On June 18, 2019, the Fifth Circuit affirmed the district court's judgment. *Id.* Lastly, the Supreme Court of the United States denied Defendant's petition for a writ of certiorari on March 23, 2020. *Jones v. Davis*, 140 S. Ct. 2519 (2020).

IT IS THEREFORE EVIDENT that Defendant has exhausted his avenues for relief through the state and federal courts, and further there are no stays of execution in effect in this case.

ACCORDINGLY, IT IS HEREBY ORDERED that the Defendant, Quintin Phillippe Jones, who has been adjudged to be guilty of capital murder as charged in the indictment and whose punishment has been assessed by the verdict of the jury and judgment of the Court at **DEATH**, shall be kept or taken into the custody of the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice until the **19th day of May 2021**, upon which day, at the Correctional Institutions Division of the Texas Department of Criminal Justice, at some time after the hour of six o'clock p.m., in a room designated by the Correctional Institutions Division of the Texas Department of Criminal Justice and arranged for the purpose of execution, the said Director, acting by and through the executioner designated by said Director, as provided by law, is hereby commanded, ordered and directed to carry out this sentence of death by intravenous injection of a substance or substances in a lethal quantity sufficient to cause the death of the Defendant, Quintin Phillippe Jones, until

Quintin Phillippe Jones is dead. Such procedure shall be determined and supervised by the said Director of the Correctional Institutions Division of the Texas Department of Criminal Justice.

IT IS FURTHER ORDERED that the **Clerk of this Court** shall issue and deliver to the **Sheriff of Tarrant County, Texas**, a **Death Warrant** in accordance with this sentence and Order, directed to the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice, at Huntsville, Texas, commanding the said Director, to put into execution the Judgment of Death against the said Quintin Phillippe Jones.

The Sheriff of Tarrant County, Texas IS HEREBY ORDERED, upon receipt of said Death Warrant, to deliver said Warrant to the Director of the Correctional Institutions Division of the Texas Department of Criminal Justice, Huntsville, Texas together with Defendant, Quintin Phillippe Jones.

IT IS FURTHER ORDERED that the **Clerk of this Court** shall deliver a copy of this order, by first-class mail, e-mail, or fax not later than the second business day after the Court enters the order, to:

1. The attorney who represented the Defendant in the most recently concluded stage of a state or federal post-conviction proceeding, Michael Mowla, michael@mowlalaw.com. *See* TEX. CODE CRIM. PROC. art. 43.141(b-1).
2. Benjamin Wolff, Director, Office of Capital and Forensic Writs, Benjamin.Wolff@ocfw.texas.gov. *See Id.*
3. Rachel Patton, Assistant Attorney General,

Rachel.Patton@oag.texas.gov.

4. The Post-Conviction Section of the Tarrant County Criminal District Attorney's Office, ccaappellatealerts@tarrantcountytx.gov.

SIGNED AND ENTERED this 18th day of November 2020.

HON. ELIZABETH BEACH
JUDGE PRESIDING
CRIMINAL DISTRICT COURT NO. 1

TEXAS COURT OF CRIMINAL APPEALS

Austin, Texas

CORRECTED MANDATE

M A N D A T E

THE STATE OF TEXAS,

TO THE CRIMINAL DISTRICT COURT NUMBER ONE OF TARRANT COUNTY — GREETINGS:

Before our **COURT OF CRIMINAL APPEALS**, on the 5th day of **NOVEMBER**, A.D. 2003, the cause upon appeal to revise or reverse your Judgment between:

QUINTIN PHILLIPPE JONES

VS.

THE STATE OF TEXAS

CCRA NO. 74,060

TRIAL COURT NO. 0744493D

was determined; and therein our said **COURT OF CRIMINAL APPEALS** made its order in these words:

"This cause came on to be heard on the record of the Court below, and the same being considered, because it is the Opinion of this Court that there was no error in the judgment, it is **ORDERED, ADJUDGED AND DECREED** by the Court that the judgment be **AFFIRMED**, in accordance with the Opinion of this Court, and that this Decision be certified below for observance."

WHEREFORE, We command you to observe the Order of our said **COURT OF CRIMINAL APPEALS** in this behalf and in all things have it duly recognized, obeyed and executed.

WITNESS, **THE HONORABLE SHARON KELLER**,

Presiding Judge of our said **COURT OF CRIMINAL APPEALS**,

with the Seal thereof annexed, at the City of Austin,

this 1st day of **DECEMBER**, A.D. 2003.

TROY C. BENNETT, JR., Clerk

Veronica Arellano, Deputy Clerk

Veronica Arellano

AFFIDAVIT OF JOHN F. EDENS, PH.D.

I, John F. Edens, swear under penalties of perjury that the information in this affidavit is true and correct:

I. Background and Qualifications

1. I have personal knowledge of the facts stated in this affidavit. I am signing this affidavit knowingly, voluntarily, and freely. I fully understand the contents of this affidavit. I read, write, and speak English.

2. I am a Full Professor in the Department of Psychological and Brain Sciences at Texas A&M University (TAMU). I am also formerly the Director of Clinical Training (2012-2016) of the doctoral training program in Clinical Psychology at TAMU, as well as a licensed psychologist for 20 years in the state of Texas until I retired my license (in good standing) in 2019. Over the years, I have been actively involved in education and training in the areas of psychological and personality assessment, violence risk assessment, forensic psychology, abnormal psychology, research methodology, and professional ethics. I have taught scores of courses within these areas to hundreds of doctoral students and thousands of undergraduate students. I also have been invited to conduct numerous advanced training workshops in these and related areas for mental health, legal, and criminal justice professionals throughout North America, Europe, Asia, and Australia.

3. I have conducted research on psychological assessment and diagnosis and the prediction of human behavior since the 1990s and have published approximately 200 peer-reviewed journal articles, book chapters, and professional manuals related to these topics. Most of my research has focused on forensic and correctional mental health assessment issues, such as the scientific reliability and validity of psychological testing and diagnosis among criminal offender populations and the potential for criminal offenders to engage in future violence and other forms of socially deviant behavior inside and outside institutional settings.[i] For example, I was a co-investigator on a $1.3 million multi-site federal research grant from the *National Institute of Mental Health* that examined the role of psychopathic personality disorder (psychopathy) and antisocial personality disorder (ASPD) in the adjustment and future conduct of criminal offenders.

4. I believe it is fair to say that my research in the area of forensic and clinical psychology has been highly influential in the scientific and professional community. For example, I am in the top 1% of cited researchers in the fields of Psychology and Psychiatry (as documented by *Essential Science Indicators*) and I have received national awards and honors from various professional and scientific organizations over the course of my career (e.g., the *Saleem Shah Award for Early Career Contributions to Law and Psychology*, jointly awarded by the American Psychology-Law Society and the American Academy of Forensic Psychology

[2001], the *Theodore Millon Award in Personality Psychology*, jointly awarded by the American Psychological Foundation and the Society of Clinical Psychology [2015]). I have also been awarded Fellow status by the two largest professional organizations in psychology in the United States: the American Psychological Association and the Association for Psychological Science.

5. I am the lead author of the *Personality Assessment Inventory Interpretive Report for Correctional Settings* (PAI-CS).[ii] The PAI-CS is an empirically derived, actuarial interpretative system designed to aid in the identification of inmates who have mental health problems and/or are likely to have difficulties adjusting to prison. This interpretive report is used in numerous state prison systems as part of their mental health screening and assessment procedures for newly incarcerated inmates.

6. I have published extensively on controversies concerning various psychiatric diagnoses, psychological tests, and assessment instruments and procedures used in forensic and correctional settings, particularly those intended to assess psychopathy, such as the Hare Psychopathy Checklist-Revised (PCL-R), as well as antisocial personality disorder (ASPD).[iii] I also have consulted with numerous prosecution offices, defense counsel, and state agencies (e.g., probation departments) on issues related to forensic mental health assessment, particularly in

terms of the scientific reliability and validity of various tests, psychiatric diagnoses, and assessment methodologies.

7. Because of my background and expertise in forensic and correctional psychology, I am frequently called on to evaluate the work of other social scientists and mental health professionals. For example, I am formerly an Associate Editor of the peer-reviewed scientific journals *Psychological Assessment*, the *Journal of Personality Assessment,* and *Assessment*. In these editorial roles, I have been responsible for judging the scientific merit of research manuscripts submitted for publication and making editorial decisions, with input from peer reviewers, regarding whether these research reports are scientifically rigorous and warrant publication. At these journals, I have been primarily responsible for evaluating submissions that focus on forensic mental health topics (e.g., psychopathic and antisocial personality disorder, malingered mental illness, violence risk assessment, adjudicative competence). I also have served on the editorial boards of multiple peer-reviewed psychology-law journals (e.g., *Law and Human Behavior*, *Behavioral Science and the Law, International Journal of Forensic Mental Health*), where I provide peer reviews for research manuscripts submitted for publication. In this capacity, I provide the Editor or Associate Editor with a review of the methodological rigor of the research and a recommendation concerning its overall contribution to the scientific literature. Over the course of

my career, I have been asked to serve as an editor or reviewer for hundreds of scientific research reports from a multitude of social science and medical journals.

8. As I noted above, I have contributed extensively to and am very familiar with the research literature on the Hare PCL-R, psychopathy, ASPD, and other personality disorders. Because of my expertise in this area, I have been asked to submit affidavits and declarations (similar in content to this document) that have expressed my grave reservations about the use of the PCL-R, labels such as "psychopath," and diagnoses of ASPD in numerous state and federal capital murder cases.

II. Referral Question

9. I was asked by defense counsel for Quintin P. Jones to review evidence presented by mental health experts who testified at his sentencing hearing in February 2001 and comment on the potential implications of the introduction of the PCL-R in Mr. Jones's capital murder trial. The PCL-R is a 20-item checklist/rating scale that is intended to be used by trained professionals to measure the personality disorder of psychopathy. The 20 items consist of prototypically psychopathic traits (e.g., remorselessness, grandiosity, superficial charm) but also include items that focus on a history of antisocial and criminal acts (e.g., juvenile delinquency, past revocation of conditional release). The PCL-R typically is scored based on a semi-structured interview and review of available

collateral information (e.g., institutional files, past mental health evaluations). Examinees can be given a score ranging from 0 (zero) to 40, with higher scores indicating that they are being rated by an examiner as more psychopathic.

10. I should highlight that I have not conducted a PCL-R evaluation of Mr. Jones and I have no opinion as to what would have been an accurate score on the PCL-R at the time of his sentencing hearing. That I have not evaluated Mr. Jones myself has no bearing on the points of concern that I raise about PCL-R evidence in this affidavit. In fact, one of the primary criticisms of this checklist in the scientific literature is that the scores derived from it in adversarial legal cases are so unreliable across different examiners that they lack any substantive probative value. Additionally, this general problem with the unreliability of PCL-R scores is evident in the competing forensic evaluations performed on Mr. Jones at the time of his original trial.

11. In his testimony describing his forensic mental health evaluation of Mr. Jones during his sentencing hearing, Dr. Randall Price provided a PCL-R total score of 31, which would place Mr. Jones at approximately the 88th percentile compared to the PCL-R's male prisoner normative sample. During the punishment phase Dr. Price diagnosed Mr. Jones as a "psychopath," stating to the jury, "A psychopath is a personality disorder that is characterized by a set of traits and behaviors that are, in a nutshell, the person doesn't have a conscious or has little

conscience." (*See* Record, Volume 36, page 58). Dr. Price also related psychopathy to a propensity for future dangerousness within the context of the first special issue. (*See* Record, Volume 36, page 74). However, another forensic mental health expert, Dr. Raymond Finn, testified during this sentencing hearing that his scoring of Mr. Jones on the PCL-R was only 9.5, which would place Mr. Jones between the 8th and 9th percentile when compared to the PCL-R's normative sample, essentially concluding that Mr. Jones would be one of the least psychopathic individuals housed in a prison environment. It is self-evident that two scores ranging from the 8th or 9th to the 88th percentile in a given case clearly reflect extreme disagreement on exactly how psychopathic Mr. Jones actually was at that time.

12. The extreme scoring discrepancies on the PCL-R that are evident in the competing evaluations of Mr. Jones unfortunately are not unique to his particular case. Although I am unfamiliar with any prior PCL-R testimony provided by Dr. Finn in other cases, I was retained as an expert witness in a recent Texas capital murder trial in which Dr. Price had administered the PCL-R to the defendant. In this case Dr. Price's score was vastly higher (placing the defendant at the 91st percentile) than a score that a TDCJ-employed mental health professional provided for the same defendant (18th percentile).

III. Relevant Scientific Literature

13. The extent to which separate mental health examiners will produce approximately similar scores for the same defendant (which in the diagnostic and testing literature is described by the term "inter-rater reliability") is not a question that appears to be commonly raised in adversarial legal settings in which PCL-R scores have been introduced.[iv] This is unfortunate because the extant scientific research indicates that PCL-R scores are highly *unreliable* in real world legal cases (as opposed to controlled scientific research studies). Several "field studies" of the PCL-R have reported that in adversarial settings, mental health experts disagree considerably on the scoring of this rating scale and, not surprisingly, results also suggest that prosecution-retained experts tend to give higher scores than do defense-retained experts.[v] It is unclear whether prosecution witnesses overestimate psychopathy, defense witnesses underestimate psychopathy, or both, but the key point is that how psychopathic defendants are described to be at trial is to some extent contingent on which side is retaining the expert witness.

14. That being said, even examiners who are employed or retained by the same "side" of a case (and examiners who are independently appointed) may give markedly different scores on the PCL-R, indicating that the scores themselves are to some extent a function of the expert conducting the assessment rather than simply being an objective assessment of the "true" level of psychopathic traits

exhibited by the defendant. More specifically, it has been estimated that over 30% of the variability in PCL-R scoring across contested legal cases is explained by the individual examiners who are conducting the evaluation rather than a reflection of genuine differences in the defendants who are being assessed. Put somewhat more simply, approximately a third of any given PCL-R score in these cases does not represent his or her actual level of psychopathic traits but instead reflects the idiosyncratic scoring approach of the person performing the evaluation—regardless of whether the expert examiner was retained by the prosecution or the defense.[vi]

15. Also of particular concern is that since the publication of the first PCL-R professional manual in 1991, it has been known that the "personality" items contained within the PCL-R (e.g., lack of remorse, inflated self-worth, conning/manipulative) have lower levels of inter-rater reliability than do the more criminogenic items (e.g., juvenile delinquency, revocation of conditional release). The more recent field studies cited above also demonstrate that personality characteristics appear to be extremely difficult to assess reliably in adversarial legal settings—which is particularly troubling given that they seem to have the most pronounced prejudicial effect on jurors[vii] (an issue to which I return in subsequent paragraphs below). Levels of inter-rater agreement in the published

field studies have been well below accepted standards of what would constitute minimal reliability for forensic mental health practice.[viii]

16. The reasons for the unreliability of psychopathy evaluations across examiners have not been fully articulated in the literature, but there is recent evidence that even those trained by the instrument developer, Robert Hare (through his Darkstone Research Group workshops), struggle to assess reliably the personality traits included in the PCL-R. Blais, Forth, and Hare (2017)[ix] summarized reliability statistics for 280 participants in this training program who went on to score a series of practice cases that were then evaluated for accuracy. The interpretation of what constitutes minimally acceptable reliability is open to some degree of interpretation, but the effects of this formalized training program on inter-rater reliability were disappointing regardless of the standard. In particular, the inter-rater reliability of the 'personality' items on the PCL-R was quite poor, indicating a large degree of variability in rating traits such as remorselessness, superficial charm, and lack of empathy. Again, it should be stressed that this unacceptable level of inter-rater reliability in assessing these personality traits was produced by professionals who had *just completed a formalized training program conducted by the developer of the instrument*, leading the authors to conclude that those raters' PCL-R scores "did not meet the standard recommended for criminal cases" (p. 762).

17. Even if PCL-R scores could be reliably produced in adversarial legal cases, there are additional reasons to question their relevance and probative value in capital murder trials. For example, mock jury research has shown that individuals who believe a defendant is highly psychopathic also believe that such a defendant will be highly dangerous in the future.[x] Despite this intuitive association between psychopathy and violence, at present there is little evidence to support the assertion that psychopathy diagnoses have any bearing on a convicted capital defendant's potential for future violent acts. That is, the available scientific studies suggest that psychopathy diagnoses are *at best very weakly related to violent behavior in U.S. prisons*. This assertion is based on the results of a published meta-analysis[xi] in which my colleagues and I statistically aggregated the results of all available individual research studies examining the relationship between the most widely used assessment of psychopathy in forensic settings, the PCL-R, and violence in U.S. prisons, which consisted of an aggregated sample size of over 800 inmates across five individual research studies.

18. Although I am very familiar with the professional literature concerning psychopathy, I am unaware of any published studies of the PCL-R that have examined whether they can reliably predict the violent behavior specifically of capital defendants, life sentenced offenders, or those offenders who are placed in administrative segregation—but the fact that PCL-R scores have not performed

well in the existing prison studies of non-capital prison inmates suggests that their very poor accuracy would be similar or worse among capital defendants serving life sentences.[xii]

19. As such, although well-controlled research studies suggest that PCL-R scores may be modestly to moderately related to future criminal behavior among individuals if they are *released back into the community*,[xiii] the available scientific findings do not support the argument that this instrument can identify prisoners who are likely to engage in serious violence while spending the rest of their lives incarcerated. Therefore, claims that an inmate is more likely to be violent in the future if serving out a life sentence because that inmate has been judged by a mental health professional to be "psychopathic" are based on almost no scientific support and actually ignore what are known to be legitimate correlates of violence in prison settings (e.g., young age, limited education, prison gang membership).

20. To the extent that PCL-R scores have a modest to moderate predictive relationship with violence if prisoners are released back into the community, it should be noted that extant research findings indicate that it is *not* the personality traits (e.g., remorselessness, conning/manipulative) related to this diagnosis that are relevant to identifying those most at risk for future violent crime. Rather, it is the more criminalistic characteristics measured by the PCL-R (e.g., juvenile delinquency, past failure on conditional release, poor behavioral controls) that are

most important to predicting criminal recidivism. Knowing whether a soon-to-be released inmate appears to lack remorse and is grandiose and unempathic is much less informative about his or her potential for future community violence than knowing whether he or she has an extensive history of irresponsible, impulsive, and criminal behavior. As such, the PCL-R items that are likely to be the *most* influential on jurors' decisions concerning a death sentence are the ones that are the *least* relevant to predicting future crime in the community.[xiv]

21. To summarize, in the context of a capital murder trial, testimony that a defendant is "a psychopath" based on a high PCL-R score is unreliable, unscientific, and misleading in relation to the likelihood of a defendant being a future danger to society if serving a life sentence in prison. Given our concerns about misuses and abuses of psychopathy evidence in capital cases, several other forensic mental health experts and I have recently detailed several of the limitations of this checklist for this specific purpose in a *Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist—Revised in Capital Sentencing to Assess Risk for Institutional Violence*.[xv]

22. In addition to having very limited probative value due to poor inter-rater reliability and almost no predictive validity for prison violence, the introduction of the PCL-R into capital proceedings has a strong likelihood of unduly prejudicing jurors against a defendant. Among venirepersons, the

psychopath label evokes images of real-world serial killers such as Ted Bundy, as well as fictional villains such as Hannibal Lecter.[xvi] In fact, among a sample of over 400 venirepersons participating in a survey in Dallas County, Texas, Charles Manson was the most common response (20%) when participants were asked to spontaneously identify the person they first thought of when hearing the term "psychopath" (followed by Jeffrey Dahmer [14%], Adolf Hitler [12%], and Ted Bundy [11%]). [xvii] In a recent meta-analysis[xviii] published by my research laboratory that examined how perceptions of psychopathic traits are related to attitudes about criminal defendants, we found that mock jurors who believe a defendant to be highly psychopathic are more likely to support death verdicts (in capital murder trial simulations) and more likely to recommend longer criminal sentences (in non-capital trial simulations) than are participants who believe a defendant to be less psychopathic. Additionally, they are more likely to rate a defendant as more dangerous and more "evil" than are participants who believe a defendant to be less psychopathic.

23. Research summarized in the preceding paragraph indicates that mock jurors who perceive a defendant to be highly psychopathic also have more punitive attitudes about their case dispositions. Such findings do not directly examine, however, the extent to which expert testimony concerning psychopathy may influence case outcomes (e.g., jury verdicts). In a series of research studies,[xix] my

colleagues and I have experimentally manipulated the presence of psychopathy evidence in capital case vignettes presented to mock jurors. The results of these studies indicate that defendants who were described as psychopaths were viewed as considerably more dangerous than defendants who were not described as psychopaths, even though all other facts of the cases other than diagnoses were described identically. In these studies, support for executing a psychopathic defendant was considerably higher than support for executing him when not described as psychopathic. For example, in one of these studies,[xx] 60% of the participants learning that the defendant was described as psychopathic indicated they would support a death sentence for the defendant, whereas only 38% did so when he was described as non-mentally disordered, and only 30% did so when he was described as psychotic (e.g., experiencing delusions and hallucinations). (In the lone study my research lab has published in which psychopathy evidence did not predict greater support for death verdicts,[xxi] post-testing of the research participants indicated that many did not understand the complicated sentencing instructions we provided them, such as the definition of mitigating evidence.) A recent meta-analysis[xxii] of this area of scientific literature confirmed that the introduction of evidence that a defendant is psychopathic in mock jury trials results in more punitive outcomes when compared to cases in which this diagnosis is not introduced.

24. I should note that some experimental research[xxiii] has suggested that expert testimony that a defendant is psychopathic may not *always* have a significant impact on mock jurors. These types of findings only seem to occur, however, when mock jurors *already* believe a defendant is highly psychopathic *prior* to reviewing any mental health evidence about his diagnostic status. In replicating some of this earlier research, my research lab[xxiv] recently demonstrated that when jurors are informed that a defendant has a history of being remorseless, manipulative, and superficial, they tend believe that he is highly psychopathic – regardless of whatever subsequent diagnostic label an expert witness provides (e.g., "psychopathic," "schizophrenic"). The results of this research suggest that, once a juror believes that a defendant is highly psychopathic, the introduction of the label "psychopath" by an expert witness may in fact have little *additional* prejudicial impact. It does *not*, however, support the conclusion that testimony about psychopathy will have little or no prejudicial impact on jurors who have yet to form an opinion about a defendant's mental health status.

25. Although mock jury studies in isolation are not dispositive in terms of establishing the stigmatizing effects of a psychopathy diagnosis, field research also has demonstrated that perceived psychopathic traits have a strong relationship with juror attitudes about criminal defendants. For example, Sundby (1998) published research from the Capital Jury Project indicating that actual jurors in capital

murder trials described defendants whom they had sentenced to death with phrases such as "blasé," "cocky," "very unremorseful," "cocksure," "nonchalant," "no remorse—almost a cocky attitude," and "clever, smart, [and] calculating."[xxv]

IV. Opinion

26. PCL-R psychopathy evidence provided by examiners in adversarial legal settings is highly unreliable, has little or no probative value concerning prison violence risk, and has the strong potential to stigmatize capital defendants with an irrelevant and pejorative label and associated set of personality traits (e.g., remorselessness, conning/manipulative). As such, it is very difficult if not impossible to argue that labeling a defendant as psychopathic has any demonstrated probative value in capital cases. As was highlighted in earlier sections of this affidavit, the general unreliability of PCL-R scores is in fact evident in this particular case, with Dr. Finn and Dr. Price providing extremely divergent scores (8^{th} or 9^{th} percentile vs 88^{th} percentile). In sum, testimony based on the PCL-R that a defendant is "a psychopath" is unreliable, unscientific, and misleading in relation to the likelihood of a defendant being a future danger to society if serving a life sentence in prison.

End of testimony.

John F. Edens, Ph.D.
Professor
Department of Psychological and Brain Sciences
Texas A&M University

On this day, April 19, 2021 (date), appeared before me, the undersigned authority, the affiant, who being duly sworn stated under oath that the above affidavit signed by the affiant is true and correct and within his or her personal knowledge.

Notary Public

LEAH MULLETT
Notary Public, State of Texas
Notary ID #12460806-8
My Commission Expires 08-26-2024

End Notes

[i] See, e.g: Buffington-Vollum, J. K., Edens, J. F., Johnson, D. W., & Johnson, J. (2002). Psychopathy as a predictor of institutional misbehavior among sex offenders: A prospective replication. *Criminal Justice and Behavior, 29,* 497-511.

Douglas, K. S., Vincent, G., & Edens, J. F. (2006). Risk for criminal recidivism: The role of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp. 533-554). New York: Guilford.

Edens, J. F., Buffington-Vollum, J. K., Colwell, K. W., Johnson, D. W., & Johnson, J. (2002). Psychopathy and institutional misbehavior among incarcerated sex offenders: A comparison of the Psychopathy Checklist-Revised and the Personality Assessment Inventory. *International Journal of Forensic Mental Health, 1,* 49-58.

Edens, J. F., & Campbell, J. S. (2007). Identifying youths at risk for institutional misconduct: A meta-analytic investigation of the Psychopathy Checklist measures. *Psychological Services, 4,* 13-27.

Edens, J. F., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law and Human Behavior, 31,* 53-75.

Edens, J. F., & Cahill, M. A. (2007). Psychopathy in adolescence and criminal recidivism in young adulthood: Longitudinal results from a multi-ethnic sample of youthful offenders. *Assessment, 14,* 57-64.

Edens, J. F., Poythress, N. G., & Lilienfeld, S. O. (1999). Identifying inmates at risk for disciplinary infractions: A comparison of two measures of psychopathy. *Behavioral Sciences and the Law, 17,* 435-443.

Edens, J. F., & Ruiz, M. A. (2006). On the validity of validity scales: The importance of defensive responding in the prediction of institutional misconduct. *Psychological Assessment, 18,* 220-224.

Edens, J. F., Skeem, J. L., & Douglas, K. S. (2006). Incremental validity analyses of the Violence Risk Appraisal Guide and the Psychopathy Checklist: Screening Version in a civil psychiatric sample. *Assessment, 13,* 368-374.

[ii] Edens, J. F., Ruiz, M. A., & PAR Staff. (2005). *Personality Assessment Inventory Interpretive Report for Correctional Settings (PAI-CS).* Odessa, FL: Psychological Assessment Resources, Inc.

[iii] See, e.g: Amenta, A. E., Guy, L. S., & Edens, J. F. (2003). Sex offender risk assessment: A cautionary note regarding measures attempting to quantify violence risk. *Journal of Forensic Psychology Practice, 3,* 39-50.

DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S...Reidy, T. J. (2020). Statement of concerned experts on the use of the Hare Psychopathy Checklist-Revised in capital sentencing to assess risk for institutional violence. *Psychology, Public Policy, and Law, 26,* 133-144.

DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S...Reidy, T. J. (2020). Death is different: Reply to Olver et al. (2020). *Psychology, Public Policy, and Law, 26,* 511-518.

Edens, J. F. (2006). Unresolved controversies concerning psychopathy: Implications for clinical and forensic decision-making. *Professional Psychology: Research and Practice, 37,* 59-65.

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment 'out of the lab' and into 'the real world': Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29,* 710-719.

Edens, J. F., Buffington-Vollum, J. K., Keilen, A., Roskamp, P., & Anthony, C. (2005). Predictions of future dangerousness in capital murder trials: Is it time to "disinvent the wheel?" *Law and Human Behavior, 29,* 55-86.

Edens, J. F., & Douglas, K. S. (2006). Assessment of interpersonal aggression and violence: Introduction to the special issue. *Assessment, 13,* 1-6.

Edens, J. F., Magyar, M. S., & Cox, J. (2013). Taking psychopathy measures "out of the lab" and into the legal system: Some practical concerns. In K. Kiehl & W. Sinnott-Armstrong (Eds.), *Handbook on psychopathy and law* (pp. 250-272). New York: Oxford University Press.

Edens, J. F., & Otto, R. K. (2001). Release decision making and planning. In J. B. Ashford, B. D. Sales, & W. H. Reid (Eds.), *Treating adult and juvenile offenders with special needs* (pp. 335-371). Washington, DC: American Psychological Association.

Edens, J. F., Petrila, J., & Kelley, S.E. (2018). Legal and ethical issues in the assessment and treatment of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (2nd ed.). New York: Guilford.

Edens, J. F., Petrila, J., & Buffington-Vollum, J. K. (2001). Psychopathy and the death penalty: Can the Psychopathy Checklist-Revised identify offenders who represent "a continuing threat to society?" *Journal of Psychiatry and Law, 29,* 433-481.

Edens, J. F., Skeem, J. L., Cruise, K. R., & Cauffman, E. (2001). Assessment of 'juvenile psychopathy' and its association with violence: A critical review. *Behavioral Sciences and the Law, 19,* 53-80.

Kelley, S. E., Edens, J. F., Mowle, E. N., Rulseh, A., & Penson, B. N. (2019). Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies. *Journal of Clinical Psychology, 75,* 627-643.

Truong, T. N., Kelley, S. E., & Edens, J. F. (2021). Does psychopathy influence juror decision-making in capital murder trials? "The devil is in the (methodological) details." *Criminal Justice and Behavior, 48,* 690-707.

[iv] DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., & Formon, D. (2014). The role and reliability of the Psychopathy Checklist-Revised in U.S. sexually violent predator evaluations: A case law survey. *Law and Human Behavior, 38,* 248-255.

DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., Koller, J. P., & Bersoff, B. (2014). Investigating the role of the Psychopathy Checklist-Revised in United States case law. *Psychology, Public Policy, and Law, 20,* 96-107.

Edens, J. F., Cox, J., Smith, S. T., DeMatteo, D., & Sörman, K. (2015). How reliable are Psychopathy Checklist-Revised scores in Canadian criminal trials? A case law review. *Psychological Assessment, 27,* 447-456.

Edens, J. F., Petrila, J., & Kelley, S. E. (2018). Legal and ethical issues in the assessment and treatment of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (2nd Ed.) (pp. 732-751). New York: Guilford.

[v] Boccaccini, M.T., Turner, D.T. & Murrie, D.C. (2008). Do some evaluators report consistently higher or lower scores on the PCL-R?: Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283.

DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., & Formon, D. (2014). The role and reliability of the Psychopathy Checklist-Revised in U.S. sexually violent predator evaluations: A case law survey. *Law and Human Behavior, 38,* 248-255.

Edens, J.F., Boccaccini, M.T., & Johnson, D.W. (2010). Inter-rater reliability of the PCL-R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioral features? *Behavioral Sciences and the Law, 28,* 106-119.

Edens, J. F., Cox, J., Smith, S. T., DeMatteo, D., & Sörman, K. (2015). How reliable are Psychopathy Checklist-Revised scores in Canadian criminal trials? A case law review. *Psychological Assessment, 27,* 447-456.

Jeandarme, I., Edens, J. F., Habets, P., Bruckers, L., Oei, K., & Bogaerts, S. (2017). Psychopathy Checklist— Revised field validity in prison and hospital settings. *Law and Human Behavior, 41*, 29–43.

Lloyd, C.D., Clark, H.J., & Forth, A.E. (2010). Psychopathy, expert testimony and indeterminate sentences: Exploring the relationship between Psychopathy Checklist-Revised testimony and trial outcome in Canada. *Legal and Criminological Psychology, 15,* 323-339.

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24,* 944–953.

Murrie, D.C., Boccaccini, M., Johnson, J. & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior,* 32, 352–362.

Murrie, D.C., Boccaccini, M., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15*, 19-53.

[vi] Boccaccini, M.T., Turner, D.T. & Murrie, D.C. (2008). Do some evaluators report consistently higher or lower scores on the PCL-R?: Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283.

Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist-Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38*, 315-324.

[vii] Edens, J.F., Boccaccini, M.T., & Johnson, D.W. (2010). Inter-rater reliability of the PCL-R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioral features? *Behavioral Sciences and the Law, 28*, 106-119.

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24,* 944–953.

Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist-Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38*, 315-324.

[viii] Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257-272.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

[ix] Blais, J., Forth, A., & Hare, R. (2017). Examining the interrater reliability of the Hare Psychopathy Checklist-Revised across a large sample of trained raters. *Psychological Assessment, 29*(6), 762-775.

[x] Kelley, S. E., Edens, J. F., Mowle, E. N., Rulseh, A., & Penson, B. N. (2019). Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies. *Journal of Clinical Psychology, 75*, 627-643.

[xi] Guy, L. S., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting and Clinical Psychology, 73,* 1056-1064.

[xii] Cunningham, M. D., & Sorensen, J. R. (2006). Actuarial models for assessing prison violence risk: Revisions and extensions of the Risk Assessment Scale for Prison (RASP). *Assessment, 13,* 253-265.

Cunningham, M. D., Sorensen, J. R., & Reidy, T. J. (2005). An actuarial model for assessment of prison violence risk among maximum security inmates. *Assessment, 12,* 40-49.

Edens, J. F., Buffington-Vollum, J. K., Keilen, A., Roskamp, P., & Anthony, C. (2005). Predictions of future dangerousness in capital murder trials: Is it time to "disinvent the wheel?" *Law and Human Behavior, 29,* 55-86.

Edens, J. F., Petrila, J., & Buffington-Vollum, J. K. (2001). Psychopathy and the death penalty: Can the Psychopathy Checklist-Revised identify offenders who represent "a continuing threat to society?" *Journal of Psychiatry and Law, 29,* 433-481.

Marquart, J. W., Ekland-Olson, S., & Sorensen, J. (1989). Gazing into the crystal ball: Can jurors accurately predict future dangerousness in capital cases? *Law and Society Review, 23,* 449-468.

Marquart, J. W., Ekland-Olson, S., & Sorensen, J. (1994). *The rope, the chair, and the needle: Capital punishment in Texas, 1923-1990.* Austin, TX: University of Texas Press.

Marquart, J. W., & Sorensen, J. (1988). Institutional and postrelease behavior of Furman-commuted inmates in Texas. *Criminology, 26,* 677-693.

Marquart, J. W., & Sorensen, J. (1989). A national study of the Furman-commuted inmates: Assessing the threat to society from capital offenders. *Loyola of Los Angeles Law Review, 23,* 5-28.

Sorensen, J. R., & Pilgrim, R. L. (2000). An actuarial risk assessment of violence posed by capital murder defendants. *Journal of Criminal Law & Criminology, 90,* 1251-1270.

Sorensen, J. R., & Wrinkle, R.D. (1996). No hope for parole: Disciplinary infractions among death-sentenced and life-without-parole inmates. *Criminal Justice and Behavior, 23,* 542-552.

Zamble, E. (1992). Behavior and adaptation in long-term prison inmates: Descriptive longitudinal results. *Criminal Justice and Behavior, 19,* 409-425.

[xiii] Kennealy, P., Skeem, J., Walters, G., & Camp, J. (2010). Do core interpersonal and affective traits of PCL-R psychopathy interact with antisocial behavior and disinhibition to predict violence? *Psychological Assessment, 22,* 569-580.

Singh, J., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31,* 499-513.

Yang, M., Wong, S., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136,* 740-767.

[xiv] Kennealy, P., Skeem, J., Walters, G., & Camp, J. (2010). Do core interpersonal and affective traits of PCL-R psychopathy interact with antisocial behavior and disinhibition to predict violence? *Psychological Assessment, 22,* 569-580.

See also:

Edens, J. F., Kelley, S. E., Lilienfeld, S. O., Skeem, J. L., & Douglas, K. S. (2015). DSM-5 antisocial personality disorder: Predictive validity in a prison sample. *Law and Human Behavior, 39,* 123-129.

Walters, G. D., & Knight, R. A. (2010). Antisocial personality disorder with and without antecedent childhood conduct disorder: Does it make a difference? *Journal of Personality Disorders, 24,* 258–271.

[xv] DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S...Reidy, T. J. (2020). Statement of concerned experts on the use of the Hare Psychopathy Checklist-Revised in capital sentencing to assess risk for institutional violence. *Psychology, Public Policy, and Law, 26,* 133-144.

See also:
DeMatteo, D., Hart, S. D., Heilbrun, K., Boccaccini, M. T., Cunningham, M. D., Douglas, K. S...Reidy, T. J. (2020). Death is different: Reply to Olver et al. (2020). *Psychology, Public Policy, and Law, 26*, 511-518.

For commentaries on this position statement, see:
Hare, R. D., Olver, M. E., Stockdale, K. C., Neumann, C. S., Mokros, A., Baskin-Sommers, A., Brand, E., Folino, J., Gacono, C., Gray, N. S., Kiehl, K., Knight, R., Leon-Mayer, E., Logan, M., Meloy, J. R., Roy, S., Salekin, R. T., Snowden, R. J., Thomson, N., ... Yoon, D. (2020). The PCL–R and capital sentencing: A commentary on "Death is different" DeMatteo et al (2020a). *Psychology, Public Policy, and Law, 26*, 519–522.

Olver, M. E., Stockdale, K. C., Neumann, C. S., Hare, R. D., Mokros, A.,Baskin-Sommers, A.,...Yoon, D. (2020). Reliability and validity of the Psychopathy Checklist-Revised in the assessment of risk for institutional violence: A cautionary note on DeMatteo et al. (2020). *Psychology, Public Policy, and Law, 26*, 490–510.

[xvi] Helfgott, J. B. (1997, March). *The popular conception of the psychopath: Implications for criminal justice policy and practice.* Paper presented at the 34th Annual Meeting of the Academy of Criminal Justice Sciences, Louisville, KY.

Rogers, R., Dion, K. L., & Lynett, E. (1992). Diagnostic validity of antisocial personality disorder: A prototypical analysis. *Law and Human Behavior, 16*, 677–689.

Smith, S. T., Edens, J. F., Clark, J., & Rulseh, A. (2014). "So, what *is* a psychopath?" Venireperson perceptions, beliefs and attitudes about psychopathic personality. *Law and Human Behavior, 38*, 490-500.

[xvii] Edens, J. F., Clark, J., Smith, S. T., Cox, J., & Kelley, S. (2013). Bold, smart, dangerous and evil: Perceived correlates of core psychopathic traits among jury panel members. *Personality and Mental Health, 7,* 143-153.

[xviii] Kelley, S. E., Edens, J. F., Mowle, E. N., Rulseh, A., & Penson, B. N. (2019). Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies. *Journal of Clinical Psychology, 75*, 627-643.

See, also: Truong, T. N., Kelley, S. E., & Edens, J. F. (2021). Does psychopathy influence juror decision-making in capital murder trials? "The devil is in the (methodological) details." *Criminal Justice and Behavior, 48*, 690-707.

[xix] Edens, J. F., Desforges, D. M., Fernandez, K., & Palac, C. A. (2004). Effects of psychopathy

and violence risk testimony on mock juror perceptions of dangerousness in a capital murder trial. *Psychology, Crime, and Law, 10,* 393-412.

Edens, J. F., Davis, K. M., Fernandez Smith, K., & Guy, L. S. (2013). No sympathy for the devil: Attributing psychopathic traits to capital murderers also predicts support for executing them. *Personality Disorders: Theory, Research, and Treatment, 4,* 175-181.

Edens, J. F., Guy, L. S., & Fernandez, K. (2003). Psychopathic traits predict attitudes toward a juvenile capital murderer. *Behavioral Sciences and the Law, 21,* 807-828.

[xx] Edens, J. F., Colwell, L. H., Desforges, D. M., & Fernandez, K. (2005). The impact of mental health evidence on support for capital punishment: Are defendants labeled psychopathic considered more deserving of death? *Behavioral Sciences and the Law, 23,* 603-625.

[xxi] Edens, J. F., Desforges, D. M., Fernandez, K., & Palac, C. A. (2004). Effects of psychopathy and violence risk testimony on mock juror perceptions of dangerousness in a capital murder trial. *Psychology, Crime, and Law, 10,* 393-412.

[xxii] Berryessa, C. M., & Wohlstetter, B. (2019). The psychopathic "label" and effects on punishment outcomes: A meta-analysis. *Law and Human Behavior, 43*(1), 9–25.

[xxiii] Saks, M. J., Schweitzer, N. J., Aharoni, E. and Kiehl, K. A. (2014). The impact of neuroimages in the sentencing phase of capital trials. *Journal of Empirical Legal Studies, 11,* 105–131.

[xxiv] Truong, T. N., Kelley, S. E., & Edens, J. F. (2021). Does psychopathy influence juror decision-making in capital murder trials? "The devil is in the (methodological) details." *Criminal Justice and Behavior, 48,* 690-707.

[xxv] Sundby, S.E. (1998). The capital jury and absolution: The intersection of trial strategy, remorse, and the death penalty. *Cornell Law Review, 83,* 1557-1598.

Psychology, Public Policy, and Law

Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist—Revised in Capital Sentencing to Assess Risk for Institutional Violence

David DeMatteo, Stephen D. Hart, Kirk Heilbrun, Marcus T. Boccaccini, Mark D. Cunningham, Kevin S. Douglas, Joel A. Dvoskin, John F. Edens, Laura S. Guy, Daniel C. Murrie, Randy K. Otto, Ira K. Packer, and Thomas J. Reidy

Online First Publication, January 30, 2020. http://dx.doi.org/10.1037/law0000223

American Psychological Association

Psychology, Public Policy, and Law

ISSN: 1076-8971

http://dx.doi.org/10.1037/law0000223

Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist—Revised in Capital Sentencing to Assess Risk for Institutional Violence

David DeMatteo
Drexel University

Stephen D. Hart
Simon Fraser University

Kirk Heilbrun
Drexel University

Marcus T. Boccaccini
Sam Houston State University

Mark D. Cunningham
Seattle, Washington

Kevin S. Douglas
Simon Fraser University

Joel A. Dvoskin
University of Arizona College of Medicine

John F. Edens
Texas A&M University

Laura S. Guy
Simon Fraser University

Daniel C. Murrie
University of Virginia

Randy K. Otto
University of South Florida

Ira K. Packer
University of Massachusetts Medical School

Thomas J. Reidy
Monterey, California

Psychopathy as measured by the Hare Psychopathy Checklist—Revised (PCL–R; Hare, 1991, 2003) is related to a range of rule-breaking and antisocial behaviors. Given this association, psychopathy has received considerable attention from researchers and legal professionals over the past several decades. Concerns remain, however, about using PCL–R scores to make precise and accurate predictions in certain contexts, including an individual's risk for committing serious violence in high-security custodial facilities. After a brief introduction to psychopathy and the PCL–R, we discuss capital sentencing in the United States and then summarize the empirical literature regarding the ability of PCL–R scores to predict violence, with a particular focus on the PCL–R's ability to predict serious institutional violence. As described, we believe the research demonstrates that the PCL–R cannot precisely or accurately predict

David DeMatteo, Department of Psychology & Thomas R. Kline School of Law, Drexel University; Stephen D. Hart, Department of Psychology, Simon Fraser University; Kirk Heilbrun, Department of Psychology, Drexel University; Marcus T. Boccaccini, Department of Psychology and Philosophy, Sam Houston State University; Mark D. Cunningham, Private Practice, Seattle, Washington; Kevin S. Douglas, Department of Psychology, Simon Fraser University; Joel A. Dvoskin, Department of Psychiatry, University of Arizona College of Medicine; John F. Edens, Department of Psychological & Brain Sciences, Texas A&M University; Laura S. Guy, Department of Psychology, Simon Fraser University; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia; Randy K. Otto, Department of Mental Health Law & Policy, University of South Florida; Ira K. Packer, Department of Psychiatry, University of Massachusetts Medical School; Thomas J. Reidy, Private Practice, Monterey, California.

The authors are presented in alphabetical order following the third author. The Statement presented in the Appendix of this article represents the consensus of our views as individual forensic mental health professionals; it does not necessarily reflect the views of this journal's Editorial Board and publisher, the American Psychological Association, or any other agencies or organizations with which we are affiliated or for which we work. As any statement derived by consensus reflects compromise in the choice of language, each member comprising the Group of Concerned Forensic Mental Health Professionals may a have preferred manner of expressing the findings and opinions that differs from that in the Statement in nuance or detail. We thank Kellie Wiltsie for providing research assistance for this article.

Correspondence concerning this article should be addressed to David DeMatteo, Department of Psychology, Drexel University, 3141 Chestnut Street, Stratton Suite 119, Philadelphia, PA 19104. E-mail: david.dematteo@drexel.edu

an individual's risk for committing serious violence in high-security custodial facilities. Finally, we present a Statement of Concerned Experts that summarizes our findings and opinions, concluding the PCL–R cannot and should not be used to make predictions that an individual will engage in serious institutional violence with any reasonable degree of precision or accuracy, especially when making high-stakes decisions about legal issues such as capital sentencing.

Keywords: psychopathy, Psychopathy Checklist—Revised, violence risk, institutional violence, capital sentencing

Supplemental materials: http://dx.doi.org/10.1037/law0000223.supp

There is an essential tension that underlies the study of psychopathy in forensic mental health. On one hand, it continues to garner considerable attention in the scientific and professional literature, and there is a large body of work indicating that the construct is related to a broad range of adverse behavioral outcomes, including antisocial and criminal conduct. Much of this literature focuses on the use of the Hare Psychopathy Checklist—Revised (PCL–R; Hare, 1991, 2003), a psychological rating scale that is often identified as the "gold standard" for the assessment of psychopathy. Perhaps the best summary of the literature to date is that symptoms of psychopathy, as assessed using the PCL–R, are associated with general rule-breaking and trouble-making across settings and populations. But serious concerns have been expressed about the limitations of using both research on psychopathy and scores on the PCL–R to make reliable (i.e., consistent) and accurate (i.e., valid) predictions. This is especially true when those predictions concern specific individuals, specific populations, specific antisocial acts, and specific settings or are made with the goal of assisting decisions about specific legal issues.

It may strike some people as confusing or even logically incoherent to conclude that the scientific and professional literature can, simultaneously, support the general usefulness of psychopathy as a construct but fail to support the use of psychopathy rating scales by forensic mental health professionals to make certain predictions of violence. Yet, that is exactly what we—a group of concerned forensic mental health professionals—believe to be true and exactly what motivated us to prepare the Statement of Concerned Experts ("Statement") presented in this article, which focuses on what we consider to be the inappropriate use of the PCL–R to draw conclusions about an individual's risk for committing serious violence in high-security custodial facilities. We conclude that the literature does not support the use of the PCL–R to predict serious institutional violence. Our interpretation of the research literature is that not only is there an absence of proof it can do so, but that the literature demonstrates it cannot do so precisely or accurately; that is, there is "proof of absence" of such an association. This conclusion has important real-world implications because PCL–R scores are sometimes offered in capital sentencing evaluations to draw conclusions regarding an offender's "future dangerousness" in the sense of risk for serious institutional violence. Not only do PCL–R scores lack probative value with respect to determining risk for serious institutional violence, there is compelling evidence to suggest that characterizing defendants as "psychopaths" has a substantial prejudicial impact that may make jurors more inclined to support the death penalty for them (Kelley, Edens, Mowle, Penson, & Rulseh, 2019). Quite simply, the question of whether or how much to rely on the PCL–R to assess risk for serious institutional violence may be a matter of life or death.

We begin this article with a brief introduction to the concept of psychopathy and the PCL–R. We then move on to discuss the relevance of risk for serious institutional violence to capital sentencing decisions and summarize what is known and what is not known with respect to the use of the PCL–R to make precise and accurate predictions of serious institutional violence. Finally, we present the full Statement that summarizes the available scientific literature and ends by concluding the PCL–R cannot make predictions that an individual will engage in serious institutional violence with any reasonable degree of precision or accuracy and should not be used for this purpose in capital sentencing evaluations

Psychopathy and the PCL–R

The disorder currently known as psychopathy has been recognized by various names for hundreds of years, but the conceptualization of psychopathy historically included a wide range of poorly defined and conceptually inconsistent traits (see Millon, Simonsen, & Birket-Smith, 1998). However, the publication of several seminal books and articles on psychopathy in the 1940s, including Cleckley's (1941) *The Mask of Sanity* and Karpman's (1946, 1948) description of primary psychopathy, marked a shift in our understanding of the disorder. As conceptualized by Cleckley, Karpman, and others who followed them, *psychopathy* refers to a distinct constellation of interpersonal, affective, and behavioral personality traits that are extreme and maladaptive, including egocentricity, lack of empathy, shallow affect, impulsivity, and a tendency to violate social norms (Hare & Neumann, 2009).

Since the 1980s, the construct of psychopathy often has been operationalized using instruments developed by Robert Hare and colleagues, including the Psychopathy Checklist (PCL; Hare, 1980), later revised and eventually commercially published as the Hare Psychopathy Checklist—Revised (PCL–R; Hare, 1991, 2003). The Hare scales (as they are sometimes referred to) appear to reflect the interpersonal and affective characteristics of the disorder highlighted by Cleckley (1941), Karpman (1946, 1948), and others better than do many other commonly used psychological tests and diagnostic criteria. We focus on the PCL–R, as it is the Hare scale that is most widely researched and most commonly used in practice by forensic mental health professionals around the world, and also because it formed the basis for the development of other rating scales, among them the Screening Version and Youth

Version of the PCL–R (PCL:SV and PCL:YV, respectively; Hart, Cox, & Hare, 1995; Forth, Kosson, & Hare, 2003).

The PCL–R is a 20-item symptom construct rating scale for the assessment of psychopathy in adult correctional offenders and forensic mental health patients (Hare, 2003). Each of the 20 items reflects a different (putative) feature or characteristic of psychopathy. Standard administration of the PCL–R includes a semistructured interview and a review of collateral records, and on the basis of this information evaluators rate the lifetime presence of each feature using a 3-point ordinal scale (briefly, 0 = *item does not apply to the individual*, 1 = *item applies to a certain extent*, 2 = *item applies*). Scores on individual items can be summed to form various composites, the most commonly used of which is total score, reflecting the unit-weighted sum of all 20 items. Total score ranges from 0 to 40, with higher scores indicating higher levels of psychopathy. Scores of 30 and higher are frequently considered diagnostic of psychopathy, although the PCL–R manual (Hare, 2003) makes clear that this is a cutoff of convenience. Although the general descriptive utility of this particular cutoff is supported by research (e.g., Hare, 1991), it is nevertheless arbitrary. There is no good theoretical or empirical basis for assuming that psychopathy forms a natural taxon; indeed, most research tends to support the view that psychopathy is most parsimoniously and usefully conceptualized in dimensional rather than categorical terms (e.g., Edens, Marcus, Lilienfeld, & Poythress, 2006). Also, there is remarkable heterogeneity—both potential and actual—among those at or above the PCL–R cutoff score of 30 and higher (e.g., Balsis, Busch, Wilfong, Newman, & Edens, 2017; see also Mokros et al., 2015; Poythress et al., 2010).

One strength of PCL–R scores is their high level of interrater reliability (i.e., agreement among independent evaluators with respect to PCL–R scores) reported in professional manuals and in many published research reports. Many studies have found that well-trained evaluators in controlled research contexts produce scores with high levels of interrater reliability and, consequently, a small standard error of measurement ("margin of error") with respect to the expected level of disagreement between raters (see DeMatteo, Murrie, Edens, & Lankford, 2019, for a review). The PCL–R manual (Hare, 2003) reports the following intraclass correlation coefficients (ICCs): the pooled ICC for male criminal offenders was .86 for a single rating (ICC_1) and .92 for the average of two ratings (ICC_2); ICC_1 was .88 and ICC_2 was .93 for the male forensic psychiatric patients; and ICC_1 was .94 and ICC_2 was .97 for the female criminal offenders, with these values suggesting acceptable reliability (Nunnally & Bernstein, 1994).

But interrater reliability is a property of scores obtained for a particular sample of people and in a particular context; it is not a stable property of the test itself that necessarily generalizes across samples or contexts. Research conducted over the past 10 to 15 years raises concerns about the interrater reliability of PCL–R scores made in psycholegal contexts, and several caselaw reviews have examined the interrater reliability of PCL–R scores in court cases. For example, in their review of United States sexually violent predator (SVP) cases involving use of the PCL–R, DeMatteo et al. (2014a) identified 29 cases in which the same offender was assessed with the PCL–R by two evaluators. In those 29 cases, the ICC_1 was .58, and only 41% of the score differences were within one standard error of measurement (*SEM*). Further, scores by prosecutor-retained experts were significantly higher than the scores produced by defense-retained experts; prosecution experts reported PCL–R scores of 30 or above in nearly 50% of the cases, compared with less than 10% of the same cases appraised by defense experts. In a caselaw survey that included 102 criminal cases from Canada, the single-rater ICC was .59 for all cases, with an ICC of .66 for cases involving a sexual offense and an ICC .46 for nonsexual offense cases (Edens, Cox, Smith, DeMatteo, & Sörman, 2015).

From a practical perspective, it is useful to note that if the ICC for PCL–R ratings in adversarial legal proceedings do in fact approximate .60 as suggested above, then the corresponding 95% confidence interval around an average PCL score would fall between the 11th and 89th percentiles (Edens & Boccaccini, 2017). This analysis ignores certain important qualifiers, such as the fact that the PCL–R normative data are not normally distributed and that reliability estimates are not constant across the range of possible test score results (i.e., they tend to decrease the further away an obtained score is from the mean), which may further reduce the expected agreement among raters (Cooke & Michie, 2010).

Taken together, these results reveal two things. First, there is a tendency for examiners in adversarial settings to disagree with each other to an extent that is much greater than would be expected based on the ICC values reported in the PCL–R professional manual (Hare, 2003). Second, there is a tendency for prosecution-retained evaluators to report higher PCL–R scores than do defense-retained evaluators in evaluations of the same person, made around the same time, and even when made on the same information base. This tendency for some experts to drift from more objective findings to ratings that better support the party that retained them has been termed *adversarial allegiance* (Murrie & Boccaccini, 2015). Adversarial allegiance has been examined in both field studies and controlled research. In the first field study to examine adversarial allegiance, Murrie, Boccaccini, Johnson, and Janke (2008) collected PCL–R scores assigned by petitioner-retained[1] and respondent-retained psychologists in 23 SVP cases in Texas; these cases permitted the examination of PCL–R scores that opposing evaluators assigned to the same offender. There was a large difference between PCL–R scores assigned by petitioner-retained and respondent-retained evaluators (Cohen's d = 1.03) that reflected a low level of interrater agreement across raters (ICC = .39). In 14 of the 23 cases (61%), there was a difference of more than 6.0 points between the two PCL–R total scores; given the *SEM* of roughly 3.0 points for PCL–R scores, differences of this magnitude should occur by chance in less than 5% of cases. In each case, the petitioner-retained evaluator assigned a higher score than the respondent-retained evaluator. A follow-up study that included 35 SVP cases revealed similar allegiance effects in PCL–R scoring (Murrie et al., 2009).

Although the results of field studies suggest the presence of adversarial allegiance in PCL–R scoring, the nature of field

[1] As civil proceedings, SVP hearings use slightly different terminology than criminal proceedings. The petitioner, which is the party seeking civil commitment of the offender, is roughly analogous to the prosecution in criminal proceedings, whereas the respondent is roughly analogous to the defendant in criminal proceedings.

studies does not permit alternative explanations for the observed results to be ruled out. It is possible, for example, that the observed pattern in PCL–R scoring in the field studies may be due to savvy attorneys selecting experts who are most favorable to their perspective on the case. It is also possible that the nature of caselaw reviews, which comprise only published cases, is contributing to the appearance of adversarial allegiance. In other words, contentious cases are more likely to go to trial, whereas the large majority of cases that never went to trial may have involved similar PCL–R scores by prosecution-retained and defense-retained experts. Finally, it is possible that one unreliable PCL–R score may lead the parties to reach a plea bargain instead of proceeding to trial, thereby making the case unavailable for research purposes.

Fortunately, some experimental research (which does not have the same limitations as field studies) has examined adversarial allegiance in PCL–R scoring. Murrie, Boccaccini, Guarnera, and Rufino (2013) recruited more than 100 forensic psychologists and psychiatrists under the guise of performing a forensic consultation. These forensic mental health professionals were (without their awareness) randomly assigned to either a prosecution-allegiance or defense-allegiance group. Participants met for 10 to 15 minutes with an attorney who posed as leading either a public defender service or specialized prosecution unit, and the attorney then requested that the expert score two tests, one of which was the PCL–R, based on extensive offender records. Each participant was scoring the same four case files that spanned from low risk to high risk. As hypothesized, the PCL–R scores assigned by prosecution experts and defense experts showed evidence of adversarial allegiance. On average, prosecution evaluators assigned significantly higher PCL–R scores than did defense evaluators for three of four cases, with effect sizes in the medium to large range (Cohen's *d* of .55 to .85). Follow-up analyses examined how likely it was that a randomly selected prosecution expert and a randomly selected defense expert would assign scores that were so different that they could not be explained by random measurement error. Results revealed that more than 20% of the score pairings for each case reflected a score difference that was more than twice the *SEM* in the PCL–R manual. Further, most large (≥2 *SEM*) differences were in the direction of adversarial allegiance, with the prosecution expert assigning higher scores and the defense expert assigning lower scores (Murrie et al., 2013).

Capital Sentencing in the United States and the Issue of Risk for Serious Institutional Violence

Capital sentencing is the process by which criminal offenders are sentenced to death or life in prison after being convicted of a capital offense. The Supreme Court of the United States has provided many substantive and procedural constitutional restrictions on imposing the death penalty. Among other rulings, the Supreme Court has held that the death penalty (a) cannot be mandatorily imposed (*Roberts v. Louisiana*, 1976); (b) can only be imposed for crimes involving death (*Kennedy v. Louisiana*, 2008); (c) cannot be imposed on individuals who were juveniles at the time of the offense (*Roper v. Simmons*, 2005), individuals who are intellectually disabled (*Atkins v. Virginia*, 2002), or individuals who are not competent to be executed (*Ford v. Wainwright*, 1986; *Panetti v. Quarterman*, 2007); (d) requires a jury to reach findings of fact concerning aggravating factors (*Ring v. Arizona*, 2002); and (e) must be based on individualized consideration of each crime and defendant (*Eddings v. Oklahoma*, 1982; *Lockett v. Ohio*, 1978).

In several death penalty decisions dating back to the reinstatement of capital punishment in 1976, the Supreme Court has held that the sentencing jury must be given guidance in deciding whether death is an appropriate punishment (e.g., *Gregg v. Georgia*, 1976). The guidance provided to sentencing juries takes the form of statutorily defined aggravating factors (which support the imposition of the death penalty) and a nonexhaustive statutory list of mitigating factors (which support the imposition of life in prison). Aggravating factors, which are intended to narrow the class of offenders for whom death is appropriate, pertain to the offense and offender (e.g., murdering certain classes of people, committing murder in the course of a felony, an offender's history of prior violent felonies), whereas mitigating factors can be anything that is relevant to the determination of whether death is an appropriate sentence. One aggravating factor outlined by some states is a capital defendant's risk of future danger (see DeMatteo, Murrie, Anumba, & Keesler, 2011; Fairfax-Columbo & DeMatteo, 2017).

In capital sentencing contexts, future dangerousness is the probability that an individual, absent a penalty of death, will engage in future violent behavior. Currently, of the 29 states that have the death penalty, three states (Oregon, Texas, and Virginia) explicitly require that sentencing juries consider future dangerousness as an aggravating factor, three states (Idaho, Oklahoma, Wyoming) explicitly permit consideration of future dangerousness as an aggravating factor, 12 states (Alabama, Arkansas, California, Colorado, Georgia, Kentucky, Missouri, North Carolina, Pennsylvania, South Carolina, South Dakota, Utah) permit consideration of future dangerousness as a nonstatutory aggravating factor, and six states (Florida, Indiana, Kansas, Mississippi, Ohio, Tennessee) prohibit consideration of future dangerousness as an aggravating factor, with the remaining five states (Louisiana, Montana, Nebraska, Nevada, New Hampshire) making no mention of future dangerousness in the death penalty statute. In many death penalty jurisdictions, considering risk for future violence is quite commonplace, with research suggesting that future dangerousness often plays a prominent role in capital sentencing contexts (e.g., Cunningham & Goldstein, 2003; Cunningham & Reidy, 1999; Shapiro, 2009).

When considering the role of future dangerousness in capital sentencing proceedings, it is important to frame the question properly. As noted, at the capital sentencing stage, the jury usually deliberates between sentencing the defendant to death or life in prison; in most cases, release to the community is not an option, at least in the foreseeable future and barring unforeseen circumstances.[2] Therefore, as noted by a number of researchers and scholars, questions about violence risk in capital cases primarily

[2] A few states impose capital sentences with the possibility of release or parole in the distant future. As such, in these cases, forensic mental health professionals may be asked to opine about the defendant's risk for violence if and when the offender is released to the community many years in the future. These circumstances are rare, however. Most violence risk assessments in capital sentencing proceedings focus on risk of violence in the prison context, and it is difficult to imagine a scenario in which a violence risk assessment for capital sentencing would address risk of violence in the community in the near future.

involve whether the defendant will be violent while incarcerated in a high-security correctional facility (e.g., Cunningham, 2006, 2008; DeMatteo et al., 2011; Edens, Buffington-Vollum, Keilen, Roskamp, & Anthony, 2005). Many courts have explicitly recognized that violence risk assessments in capital cases are specific to the prison context (e.g., *United States v. Sablan*, 2006), although some have taken a broad and amorphous view of what it means to be a potential "danger to society" (e.g., *Coble v. Texas*, 2010). In this article, we are focusing specifically on the use of the PCL–R to predict serious (i.e., nontrivial) violence in high-security correctional settings. It is generally accepted in the field of forensic mental health that violence risk is—and therefore violence risk assessment must be—context specific (e.g., Conroy & Murrie, 2007; Heilbrun, 1992, 2009).

As this review makes clear, future dangerousness may be a relevant consideration in capital sentencing evaluations. The PCL–R has been used to asses psychopathy as a risk factor for future violence in such evaluations (e.g., *Busby v. Stephens*, 2015; *Martinez v. Dretke*, 2004; *United States v. Barnette*, 2000; *United States v. Fell*, 2008). Accordingly, it is essential to evaluate the predictive validity of PCL–R scores, and in particular to evaluate its predictive validity with respect to serious institutional violence. In the following section, we turn to this issue.

Predictive Validity of the PCL–R

As noted previously, a well-developed body of research suggests that psychopathy is related to several outcomes that are of considerable interest to the criminal justice system. PCL–R scores are associated with diverse forms of antisocial and criminal behavior in diverse settings and populations (see Patrick, 2018, for a review). As a result, researchers, clinicians, and legal professionals are attentive to psychopathy in a variety of legal contexts. Research suggests that psychopathy evidence, typically in the form of PCL–R scores, is offered in legal proceedings in the United States, the United Kingdom, and Canada (DeMatteo et al., 2014b; Gagnon, Douglas, & DeMatteo, 2007; Howard, Khalifa, Duggan, & Lumsden, 2012). PCL–R scores may be of use in some psycholegal evaluations when considered as part of a comprehensive, individualized, and contextualized evaluation. But because of their imperfect interrater reliability (which is, of course, a concern in any evaluation) and variability in their predictive validity across outcomes, settings, and samples (which is a concern with respect to prediction of serious institutional violence), PCL–R scores may lack probative value or, worse, have a prejudicial impact. (For a fuller discussion of the potential prejudicial impact of PCL–R scores, see DeMatteo, Hodges, & Fairfax-Columbo, 2016, and DeMatteo et al., 2019.)

The predictive validity (accuracy) of PCL–R scores with respect to general institutional misconduct has been studied for many years. Some early retrospective studies provided evidence of an association between PCL–R scores and past institutional misconduct. However, to the extent that PCL–R scores could have been biased or contaminated by the violence history of people being evaluated, such research is of little or no value in evaluating predictive accuracy. Later studies specifically examined the ability of PCL measure scores to predict institutional misconduct using true prospective research designs. In meta-analyses of such studies, Walters (2003a) reported a moderate association between PCL–R Total scores and institutional adjustment, including both violent and nonviolent institutional conduct (r_w = 0.27), and small (r_w = 0.18) to moderate (r_w = .27) associations between PCL–R Factor 1 and 2 scores, respectively, for violent and nonviolent infractions (Walters, 2003b). Still, Walters (2003a, 2003b) did not distinguish between more serious institutional violence and other infractions.

In a large meta-analysis of published and unpublished studies, Guy, Edens, Anthony, and Douglas (2005) coded 273 effect sizes to examine the association between PCL, PCL–R, and PCL:SV scores and institutional misconduct in civil psychiatric, forensic psychiatric, and correctional facilities. Importantly, they were able to specifically analyze the association with more serious institutional violence, in this case, physical violence (i.e., any actual or attempted physical harm). The association between total scores and physical violence was small (r_w = .17) – indeed, much smaller than the typical violence risk assessment meta-analytic effect sizes, which are best described as moderate in size (*r*s ≅ .30–35; see Campbell, French, & Gendreau, 2009; Fazel, Singh, Doll, & Grann, 2012; Singh, Grann, & Fazel, 2011; for a review of risk assessment meta-analyses, see Douglas, 2019).

A few studies published after the Guy et al. (2005) meta-analysis found that PCL measures predict institutional misconduct (e.g., Huchzermeier, Bruss, Geiger, Kernbichler, & Aldenhoff, 2008), but most studies have reported similarly weak effects (e.g., Camp, Skeem, Barchard, Lilienfeld, & Poythress, 2013; Hogan & Olver, 2016; McDermott, Edens, Quanbeck, Busse, & Scott, 2008; Morrissey et al., 2007; Walters & Mandell, 2007). It should also be noted that the rate of serious institutional violence among capital murderers sentenced to death is very low (e.g., Cunningham, Reidy, & Sorensen, 2005; Sorensen & Wrinkle, 1996), which would tend to further reduce the predictive validity of the PCL–R.

Conclusion

Two major findings emerged from our review of the literature, summarized above. First, the interrater reliability of PCL–R scores in field settings, and in particular in adversarial contexts, is problematically low. Second, the overall association between PCL–R scores and violence at the group level is only moderate in terms of effect size, both in absolute terms and relative to the effect size of other established risk factors for violence; the association between PCL–R scores and violence in institutional settings is small in terms of effect size; and the association between PCL–R scores and serious institutional violence is negligible. Our conclusion based on these findings was that one cannot use the PCL–R in the context of capital sentencing evaluations to make predictions that an individual will engage in serious violence in high-security institutional settings with adequate precision or accuracy to justify reliance on the PCL–R scores.

Accordingly, we established a Group of Concerned Forensic Mental Health Professionals and developed a Statement to summarize our findings and opinions in this respect (see the Appendix and the online supplemental materials). Our goal in developing and disseminating the Statement was to educate others concerning the current state of the scientific literature and the appropriate use of the PCL–R when making capital sentencing and other high-stakes decisions. We emphasize that although this Statement focuses on the PCL–R, this is only because it is the instrument most widely

used to assess psychopathy in forensic mental health contexts; all our concerns about relying on the PCL–R to predict whether an individual will commit serious institutional violence apply equally or to an even greater degree to the use of other means of assessing psychopathy for that purpose.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Press.

Atkins v. Virginia, 536 U.S. 304 (2002).

Balsis, S., Busch, A. J., Wilfong, K. M., Newman, J. W., & Edens, J. F. (2017). A statistical consideration regarding the threshold of the Psychopathy Checklist—Revised. *Journal of Personality Assessment, 99,* 494–502. http://dx.doi.org/10.1080/00223891.2017.1281819

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL–R scores than others? Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283. http://dx.doi.org/10.1037/a0014523

Busby v. Stephens, 2015 WL 1037460 (N. D. Tex. Mar. 10, 2015).

Camp, J. P., Skeem, J. L., Barchard, K., Lilienfeld, S. O., & Poythress, N. G. (2013). Psychopathic predators? Getting specific about the relation between psychopathy and violence. *Journal of Consulting and Clinical Psychology, 81,* 467–480. http://dx.doi.org/10.1037/a0031349

Campbell, M. A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36,* 567–590. http://dx.doi.org/10.1177/0093854809333610

Cleckley, H. (1941). *The mask of sanity*. St. Louis, MO: Mosby.

Coble v. Texas, 330 S. W. 3d 253 (Tx. Ct. App. 2010).

Conroy, M. E., & Murrie, D. C. (2007). *Forensic assessment of violence risk: A guide for risk assessment and risk management*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781118269671

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior, 34,* 259–274. http://dx.doi.org/10.1007/s10979-009-9176-x

Cunningham, M. D. (2006). Dangerousness and death: A nexus in search of science and reason. *American Psychologist, 61,* 828–839. http://dx.doi.org/10.1037/0003-066X.61.8.828

Cunningham, M. D. (2008). Forensic psychology evaluations at capital sentencing. In R. Jackson (Ed.), *Learning forensic assessment* (pp. 211–238). New York, NY: Routledge/Taylor & Francis Group.

Cunningham, M. D., & Goldstein, A. M. (2003). Sentencing determinations in death penalty cases. In A. M. Goldstein & I. B. Weiner (Eds.), *Handbook of psychology: Vol. 11. Forensic psychology* (pp. 407–436). Hoboken, NJ: Wiley.

Cunningham, M. D., & Reidy, T. J. (1999). Don't confuse me with the facts: Common errors in violence risk assessment at capital sentencing. *Criminal Justice and Behavior, 26,* 20–43. http://dx.doi.org/10.1177/0093854899026001002

Cunningham, M. D., Reidy, T. J., & Sorensen, J. R. (2005). Is death row obsolete? A decade of mainstreaming death-sentenced inmates in Missouri. *Behavioral Sciences & the Law, 23,* 307–320. http://dx.doi.org/10.1002/bsl.608

DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., & Formon, D. (2014a). The role and reliability of the Psychopathy Checklist—Revised in U.S. sexually violent predator evaluations: A case law survey. *Law and Human Behavior, 38,* 248–255. http://dx.doi.org/10.1037/lhb0000059

DeMatteo, D., Edens, J. F., Galloway, M., Cox, J., Smith, S. T., Koller, J. P., & Bersoff, B. (2014b). Investigating the role of the Psychopathy Checklist—Revised in United States case law. *Psychology, Public Policy, and Law, 20,* 96–107. http://dx.doi.org/10.1037/a0035452

DeMatteo, D., Hodges, H., & Fairfax-Columbo, J. (2016). An examination of whether Psychopathy Checklist-Revised (PCL–R) evidence satisfies the relevance/prejudice admissibility standard. In B. H. Bornstein & M. K. Miller (Eds.), *Advances in psychology and law* (Vol. 2, pp. 205–239). New York, NY: Springer. http://dx.doi.org/10.1007/978-3-319-43083-6_7

DeMatteo, D., Murrie, D. C., Anumba, N. M., & Keesler, M. E. (2011). *Forensic mental health assessments in death penalty cases*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195385809.001.0001

DeMatteo, D., Murrie, D. C., Edens, J. F., & Lankford, C. (2019). Psychopathy in the courts. In M. DeLisi (Ed.), *Routledge international handbook of psychopathy and crime* (pp. 645–664). New York, NY: Routledge/Taylor & Francis Group.

Douglas, K. S. (2019). Evaluating and managing risk for violence using structured professional judgment. In A. Day, C. Hollin, & D. Polaschek (Eds.), *Handbook of correctional psychology* (pp. 427–445). Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/9781119139980.ch26

Eddings v. Oklahoma, 455 U.S. 104 (1982).

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment "out of the lab" and into "the real world": Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29,* 599–610. http://dx.doi.org/10.1037/pas0000475

Edens, J. F., Boccaccini, M. T., & Johnson, D. W. (2010). Inter-rater reliability of the PCL–R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioral features? *Behavioral Sciences & the Law, 28,* 106–119. http://dx.doi.org/10.1002/bsl.918

Edens, J. F., Buffington-Vollum, J. K., Keilen, A., Roskamp, P., & Anthony, C. (2005). Predictions of future dangerousness in capital murder trials: Is it time to "disinvent the wheel"? *Law and Human Behavior, 29,* 55–86. http://dx.doi.org/10.1007/s10979-005-1399-x

Edens, J. F., Cox, J., Smith, S. T., DeMatteo, D., & Sörman, K. (2015). How reliable are Psychopathy Checklist-Revised scores in Canadian criminal trials? A case law review. *Psychological Assessment, 27,* 447–456. http://dx.doi.org/10.1037/pas0000048

Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G., Jr. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology, 115,* 131–144. http://dx.doi.org/10.1037/0021-843X.115.1.131

Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review, 81,* 417–480.

Fairfax-Columbo, J., & DeMatteo, D. (2017). Reducing the dangers of future dangerousness testimony: Applying the Federal Rules of Evidence to capital sentencing. *The William and Mary Bill of Rights Journal, 25,* 1047–1072.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: Systematic review and meta-analysis. *BMJ: British Medical Journal, 345,* e4692. http://dx.doi.org/10.1136/bmj.e4692

Ford v. Wainwright, 477 U.S. 399 (1986).

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Psychopathy Checklist: Youth Version*. Toronto, Canada: Multi-Health Systems.

Gagnon, N., Douglas, K., & DeMatteo, D. (2007, June). *The introduction of the Psychopathy Checklist—Revised in Canadian courts: Uses and misuses*. Paper presented at the 7th Annual Conference of the International Association of Forensic Mental Health Services, Montreal, Quebec, Canada.

Gregg v. Georgia, 428 U.S. 153 (1976).

Guy, L. S., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting and Clinical Psychology, 73,* 1056–1064. http://dx.doi.org/10.1037/0022-006X.73.6.1056

Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences, 1,* 111–119. http://dx.doi.org/10.1016/0191-8869(80)90028-8

Hare, R. D. (1991). *The Hare Psychopathy Checklist—Revised manual*. North Tonawanda, NY: Multi-Health Systems.

Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised manual* (2nd ed.). North Tonawanda, NY: Multi-Health Systems.

Hare, R. D., & Neumann, C. S. (2009). Psychopathy: Assessment and forensic implications. *Canadian Journal of Psychiatry, 54,* 791–802. http://dx.doi.org/10.1177/070674370905401202

Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences & the Law, 31,* 81–102. http://dx.doi.org/10.1002/bsl.2049

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare PCL: Screening version*. North Tonawanda, NY: Multi-Health Systems.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *The British Journal of Psychiatry, 190*(S49), S60–S65. http://dx.doi.org/10.1192/bjp.190.5.s60

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16,* 257–272. http://dx.doi.org/10.1007/BF01044769

Heilbrun, K. (2009). *Evaluation for risk of violence in adults*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/med:psych/9780195369816.001.0001

Hogan, N. R., & Olver, M. E. (2016). Assessing risk for aggression in forensic psychiatric inpatients: An examination of five measures. *Law and Human Behavior, 40,* 233–243. http://dx.doi.org/10.1037/lhb0000179

Howard, R., Khalifa, N., Duggan, C., & Lumsden, J. (2012). Are patients deemed 'dangerous and severely personality disordered' different from other personality disordered patients detained in forensic settings? *Criminal Behaviour and Mental Health, 22,* 65–78. http://dx.doi.org/10.1002/cbm.827

Huchzermeier, C., Bruss, E., Geiger, F., Kernbichler, A., & Aldenhoff, J. (2008). Predictive validity of the psychopathy checklist: Screening version for intramural behaviour in violent offenders-a prospective study at a secure psychiatric hospital in Germany. *The Canadian Journal of Psychiatry / La Revue canadienne de psychiatrie, 53,* 384–391. http://dx.doi.org/10.1177/070674370805300608

Karpman, B. (1946). A yardstick for measuring psychopathy. *Federal Probation, 10,* 26–31.

Karpman, B. (1948). The myth of the psychopathic personality. *The American Journal of Psychiatry, 104,* 523–534. http://dx.doi.org/10.1176/ajp.104.9.523

Kelley, S. E., Edens, J. F., Mowle, E. N., Penson, B. N., & Rulseh, A. (2019). Dangerous, depraved, and death-worthy: A meta-analysis of the correlates of perceived psychopathy in jury simulation studies. *Journal of Clinical Psychology, 75,* 627–643. http://dx.doi.org/10.1002/jclp.22726

Kennedy v. Louisiana, 554 U.S. 407 (2008).

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15,* 155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Leistico, A. M. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the hare measures of psychopathy to antisocial conduct. *Law and Human Behavior, 32,* 28–45. http://dx.doi.org/10.1007/s10979-007-9096-6

Lockett v. Ohio, 438 U.S. 586 (1978).

Martinez v. Dretke, 99 Fed. Appx. 538 (5th Cir. 2004).

McDermott, B. E., Edens, J. F., Quanbeck, C. D., Busse, D., & Scott, C. L. (2008). Examining the role of static and dynamic risk factors in the prediction of inpatient violence: Variable- and person-focused analyses. *Law and Human Behavior, 32,* 325–338. http://dx.doi.org/10.1007/s10979-007-9094-8

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment, 24,* 944–953. http://dx.doi.org/10.1037/a0028411

Millon, T., Simonsen, E., & Birket-Smith, M. (1998). Historical conceptions of psychopathy in the United States and Europe. In T. Millon & E. Simonsen (Eds.), *Psychopathy: Antisocial, criminal, and violent behavior* (pp. 3–31). New York, NY: Guilford Press.

Mokros, A., Hare, R. D., Neumann, C. S., Santtila, P., Habermeyer, E., & Nitschke, J. (2015). Variants of psychopathy in adult male offenders: A latent profile analysis. *Journal of Abnormal Psychology, 124,* 372–386. http://dx.doi.org/10.1037/abn0000042

Morrissey, C., Hogue, T., Mooney, P., Allen, C., Johnston, S., Hollin, C., . . . Taylor, J. L. (2007). Predictive validity of the PCL–R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry & Psychology, 18,* 1–15. http://dx.doi.org/10.1080/08990220601116345

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among expert witnesses. *Annual Review of Law and Social Science, 11,* 37–55. http://dx.doi.org/10.1146/annurev-lawsocsci-120814-121714

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24,* 1889–1897. http://dx.doi.org/10.1177/0956797613481812

Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior, 32,* 352–362. http://dx.doi.org/10.1007/s10979-007-9097-5

Murrie, D. C., Boccaccini, M. T., Turner, D., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15,* 19–53. http://dx.doi.org/10.1037/a0014897

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Panetti v. Quarterman, 551 U.S. 930 (2007).

Patrick, C. J. (Ed.), (2018). *Handbook of psychopathy* (2nd ed.). New York, NY: Guilford Press.

Poythress, N. G., Edens, J. F., Skeem, J. L., Lilienfeld, S. O., Douglas, K. S., Frick, P. J., . . . Wang, T. (2010). Identifying subtypes among offenders with antisocial personality disorder: A cluster-analytic study. *Journal of Abnormal Psychology, 119,* 389–400. http://dx.doi.org/10.1037/a0018611

Ring v. Arizona, 536 U.S. 584 (2002).

Roberts v. Louisiana, 428 U.S. 325 (1976).

Roper v. Simmons, 543 U.S. 551 (2005).

Shapiro, M. (2009). An overdose of dangerousness: How "future dangerousness" catches the least culpable capital defendants and undermines the rationale for the executions it supports. *American Journal of Criminal Law, 35,* 145–200. http://dx.doi.org/10.2139/ssrn.1402362

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review, 31,* 499–513. http://dx.doi.org/10.1016/j.cpr.2010.11.009

Sorensen, J. R., & Wrinkle, R. D. (1996). No hope for parole: Disciplinary infractions among death-sentenced and life-without-parole inmates.

Criminal Justice and Behavior, 23, 542–552. http://dx.doi.org/10.1177/0093854896023004002

Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist—Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38,* 315–324. http://dx.doi.org/10.1037/lhb0000063

United States v. Barnette, 211 F. 3d 803 (4th Cir. 2000).

United States v. Fell, 531 F. 3d 197 (2nd Cir. 2008).

United States v. Sablan, 555 F. Supp. 2d 1177 (D. Colo, 2006).

Walters, G. D. (2003a). Predicting criminal justice outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences & the Law, 21,* 89–102. http://dx.doi.org/10.1002/bsl.519

Walters, G. D. (2003b). Predicting institutional adjustment and recidivism with the psychopathy checklist factor scores: A meta-analysis. *Law and Human Behavior, 27,* 541–558. http://dx.doi.org/10.1023/A:1025490207678

Walters, G. D., & Mandell, W. (2007). Incremental validity of the psychological inventory of criminal thinking styles and psychopathy checklist: Screening version in predicting disciplinary outcome. *Law and Human Behavior, 31,* 141–157. http://dx.doi.org/10.1007/s10979-006-9051-y

World Health Organization. (1992). *ICD-10: International statistical classification of diseases and related health problems* (10th rev.). Geneva, Switzerland: Author.

Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *The International Journal of Forensic Mental Health, 16,* 227–238. http://dx.doi.org/10.1080/14999013.2017.1317302

Appendix

Statement of Concerned Experts on the Use of the Hare Psychopathy Checklist—Revised (PCL–R) in Capital Sentencing to Assess Risk for Institutional Violence

We, a group of concerned forensic mental health professionals comprising the individuals listed in Attachment A, state the following:

1. It is our consensus opinion that the Hare Psychopathy Checklist—Revised (PCL–R), a quantitative psychological test (Hare, 1991, 2003), is not generally accepted in the field of forensic mental health as a reliable and valid means of predicting serious institutional violence, that is, of estimating or determining the likelihood that a person will commit such violence in the future.

2. Our qualifications and the foundation of our consensus opinion are set out herein.

Qualifications

3. We are forensic scientists who have helped to develop, validate, and test the PCL–R in both laboratory and real-world settings and are familiar with research and practice related to the PCL–R.

4. We are active as researchers or practitioners in the field of forensic mental health. We have played prominent roles in that field as members of scientific and professional associations or the editorial boards of leading scientific and professional journals. We have conducted research on the evaluation of the PCL–R and presented the findings of our research in the form of articles in peer-reviewed journals, books and book chapters, and conference presentations. Many of us have conducted training workshops on the clinical-forensic use of the PCL–R. Many of us have used the PCL–R in the course of our practice as forensic mental health professionals, and some of us have been qualified to give expert testimony about or based on the PCL–R before courts throughout the United States.

5. None of us has an actual, potential, or perceived conflict of interest with respect to the PCL–R by which we would gain commercially or in some other way from offering the specific opinions herein or that would otherwise compromise our neutrality or objectivity.

The Nature of Quantitative Psychological Tests

6. Psychological tests are, most generally, evaluative devices or procedures intended to provide information relevant to some target construct that is either a real object (i.e., a part of the natural world) or an ideal object (i.e., a linguistic, inferential, or theoretical concept). Some (but not all) psychological tests are quantitative in nature, relying on numeric algorithms to generate scores or decisions that measure (i.e., gauge, represent, or predict) the target construct.

7. In contemporary practice, quantitative psychological tests are developed and evaluated using psychometric theory, which is a set of concepts, principles, and statistical procedures designed specifically for that purpose.

(*Appendix continues*)

8. Two primary concepts in psychometric theory are reliability and validity. In this context, reliability is potential freedom from measurement error and reflects the degree to which test scores or decisions may be precise, replicable, stable, and consistent; and validity is potential meaningfulness of measurement and reflects the degree to which test scores or decisions may be logically or empirically coherent with, representative of, or predictive of the target construct. Reliability limits validity: Test scores or decisions may be high in reliability and low in validity (e.g., precise measures of the wrong thing) but cannot be high in validity unless they are also high in reliability.

9. The steps in developing a quantitative psychological test typically include: derivation, or selection of its format and content; initial validation (also known as construction), or administration of the test in one or more data sets with the goal of exploring the reliability and validity of test scores or decisions and refining the test's format and content; and cross-validation (also known as calibration), or confirmation of the reliability and validity of test scores or decisions made using the final version of the test in one or more new data sets.

10. In forensic mental health practice, quantitative psychological test scores and decisions are expected to have a high level of reliability and validity, due to the important potential consequence of forensic decisions. The decision to use a quantitative psychological test therefore requires evaluators to conclude that the test scores or decisions are likely to have both high reliability and high validity in the case at hand. This conclusion requires two things: First, there is a body of research that provides strong direct or indirect support of the test's reliability and validity for similar purposes, in similar contexts, and for people with similar background; and second, the evaluators have sufficient expertise (i.e., training, supervision, and experience) in the use of the test to ensure they can accurately and appropriately administer, score, and interpret the test. Use of a quantitative psychological test in the absence of supporting research or sufficient expertise is contrary to standards of practice in forensic mental health.

The PCL–R

11. The PCL–R is a specific type of quantitative psychological test known as a symptom construct rating scale. It is designed to assess features of a construct known as psychopathic personality disorder in correctional and forensic mental health settings. There is active debate in the scientific community concerning the nature of the construct of psychopathic personality disorder and how best to measure it. It is not included as a distinct diagnostic category in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–5*; American Psychiatric Association, 2013) or in the tenth edition of the *International Statistical Classification of Diseases and Related Health Problems* (ICD-10; World Health Organization, 1992). There is active debate concerning the degree to which the nature of the construct of psychopathic personality disorder and way in which it is measured using the PCL–R relate to antisocial personality disorder as defined and diagnosed according to *DSM–5* and dissocial personality disorder as defined and diagnosed in *ICD-10*.

12. The PCL–R comprises 20 individual items, presented in Attachment B. Each item is defined in detail in the test manual. Trained evaluators use judgment to rate each feature on a 3-point scale (briefly, 0 = *absent*, 1 = *partially present*, 2 = *present*) based on all available clinical data, including an interview with and observation of the person, interviews with collateral informants, and case history information.

13. Scores on the individual PCL–R items are summed to yield facet, factor, and total scores. Total scores, comprising all 20 items, are relied on most heavily as a global measure of the construct in research and practice. The PCL–R test manual suggests that total scores of 30 and higher (out of a maximum possible 40 points) are generally considered indicative of psychopathic personality disorder.

Reliability of PCL–R Scores in Forensic Mental Health Practice

14. Because the PCL–R is a symptom construct rating scale, PCL–R scores rely heavily on the judgment of evaluators. For this reason, a specific facet of reliability known as interrater reliability—that is, measurement precision related to agreement between evaluators with respect to test scores—is an issue of paramount importance. In particular, it is critical to understand how this interrater reliability impacts the expected disagreement between two independent evaluators, rating the same person at the same time on the basis of the same information, with respect to the PCL–R total scores they obtain; for the sake of simplicity, we will refer to this expected disagreement as the "margin of error" of PCL–R total scores.

(*Appendix continues*)

15. Prior to the mid-2000s, the available research evidence indicated that, overall, the interrater reliability of PCL–R scores was moderate in magnitude. But the research base at that time had two important limitations:

    a. Most studies were conducted for the purpose of research or in research settings, in which the PCL–R was administered by specially trained research assistants under conditions of anonymity; there was an absence of studies on interrater reliability conducted in the context of forensic mental health practice or in applied settings (i.e., "field settings"), in which the PCL–R was administered by health care professionals as part of routine clinical or forensic practice.

    b. Most studies used statistical methods of older rather than more contemporary psychometric theory (i.e., Classical Test Theory as opposed to Generalizability Theory and Modern Test Theory).

16. Since the mid-2000s, several studies on the interrater reliability of PCL–R scores were conducted in the context of forensic mental health practice or in applied settings, or used methods of contemporary psychometric theory. These studies have yielded two new and important findings.

17. The first new and important finding is that the interrater reliability of PCL–R scores is often substantially lower when the test is evaluated in the context of forensic mental health practice or in applied settings than it is when evaluated for research purposes or in research settings. The interrater reliability of PCL–R is typically indexed using intraclass correlation coefficients (ICCs). There are actually many different specific types of ICCs, all of which reflect the agreement between evaluators under different conditions or assumptions. ICCs have a theoretical range from −1 (*perfect disagreement among two or more evaluators*) to 0 (*chance levels of agreement*) to + 1 (*perfect agreement*). Prior to the mid-2000s, the ICCs reported for agreement between independent evaluators working in research contexts were typically summarized as falling in the range of .80 to .90 (e.g., Hare, 1991, 2003), which may be characterized according to various interpretive guidelines as "good" but not "excellent" (Koo & Li, 2016). Since that time, however, studies in field settings reported ICCs that were much lower, falling in the range of .40 to .70 (e.g., Boccaccini, Turner, & Murrie, 2008; Edens, Boccaccini, & Johnson, 2010; Sturup et al., 2014), which may be characterized as "poor" to "moderate" (Koo & Li, 2016). The relatively low interrater reliability observed in field settings can be attributed in part to the limited quality and quantity of information on which evaluators relied, as well as to the limited training, supervision, and experience of those evaluators; although there is further evidence that it may also be due to the adverse impact of adversarial proceedings on the judgment of evaluators (DeMatteo et al., 2014b; Edens et al., 2015; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Murrie et al., 2013; Murrie et al., 2008; Murrie, Boccaccini, Turner, Meeks, Woods, & Tussey, 2009). This phenomenon has been referred to as "adversarial bias" or "allegiance bias" and may be considered a special case of what is referred to more generally in forensic decision making as "confirmatory bias" (Zapf & Dror, 2017).

18. The second new and important finding is that, for a given estimate of the interrater reliability of PCL–R scores, the expected disagreement between evaluators or "margin of error" is substantially larger than was estimated previously. For example, prior to the mid2000s, the expected disagreement for PCL–R total scores was estimated to be ±3 points (out of a total of 40 points) in 68% of cases, and ±6 points in 95% of cases (e.g., Hare, 1991, 2003). Put simply, the total scores of two independent evaluators were expected to be within 3 points of each other most of the time, and within 6 points almost all the time. But since that time, more precise calculations based on contemporary psychometric theory indicate the margin of error—even assuming the same level of interrater reliability, that is, .85—is actually ±3 points in only 68% of cases, but ±9 points in 95% of cases (e.g., Cooke & Michie, 2010). Additional analyses indicate that even this is an overly optimistic estimate of the margin of error, for two reasons (Cooke & Michie, 2010). First, it assumes that the interrater reliability of PCL–R total scores is about .85, whereas in field settings the interrater reliability may be considerably lower. Second, it is an estimate of the margin of error around the center of the distribution of PCL–R scores (i.e., about 20 points out of 40); however, the margin of error in fact becomes asymmetric and increases as scores approach the extremes or "tails" of the distribution (i.e., <10 and >30). This means the margin of error is larger at or around the score typically used to define psychopathic personality disorder, which is 30 points or higher out of 40. Thus, if one assumes that the interrater reliability of PCL–R scores is .80 (i.e., only slightly lower than the value of .85 assumed in the PCL–R manual), and assuming the evaluator reported a PCL–R total score of 30 points out of 40, then the total score obtained by independent evaluators would be expected to fall somewhere between 24 and 33 points out of 40 in 68% of cases, and between 19 and 36 points in 95% of cases (Cooke & Michie, 2010). In sum, the consequence of this large margin of error is considerable—and possibly even grave—uncertainty about the accuracy of a PCL–R total score obtained by a given evaluator. For example, if an evaluator administers the PCL–R and obtains a total score of 30, then one out of three evaluators who independently readministered the PCL–R would obtain scores less than or equal to 23 or, alternatively, greater than or equal to 34. This is true even assuming the interrater reliability for PCL–R total scores is good (i.e., .80), the evaluators all have the same level of training and experience, and the assessments were conducted at the same time and on the basis of the same information.

(*Appendix continues*)

Validity of PCL–R With Respect to Prediction of Serious Institutional Violence

19. According to the test manual and the writings of the test developer, the PCL–R was not developed and is not recommended to estimate the likelihood or predict that a person will commit violence in the future, either in the community or in an institution. As the test manual states, "Properly used, the PCL–R provides a reliable and valid assessment of an important clinical construct—psychopathy. **Strictly speaking, that is all it does**" (Hare, 2003, p. 15; emphasis in original).

20. Prior to the mid-2000s, the available research evidence indicated that, overall, PCL–R scores were associated with increased risk for violence in general; but they could not be used, either on their own or in combination with other risk factors, to estimate the likelihood of or predict future institutional violence by an individual with high reliability or validity. There were at least two major reasons for this:

    a. There was little or no research on the prediction of serious institutional violence using the PCL–R generally, and none at all on the prediction of serious violence in federal prisons in the United States.

    b. There was no research at all on the prediction of violence using the PCL–R at the individual level, as opposed to the group level.

21. Since the mid-2000s, several studies on prediction of serious institutional violence using the PCL–R have been conducted. These studies have yielded two new and important findings.

22. The first new and important finding is that the predictive validity of PCL–R scores is inadequate to support its use as a tool to assess risk for serious institutional violence. For example, a number of meta-analytic reviews of the literature (e.g., Campbell et al., 2009; Guy et al., 2005; Leistico, Salekin, DeCoster, & Rogers, 2008) have demonstrated that the association between PCL–R total scores and serious institutional violence is limited; and, furthermore, the magnitude of the association tended to be even smaller in studies that were conducted in prisons (as opposed to forensic mental health facilities) or in the United States (as opposed to other countries).

23. The second new and important finding is that there are significant challenges inferring an individual's likelihood of recidivism from group-level data with a high degree of accuracy and precision. A number of scholars (e.g., Cooke & Michie, 2010; Faigman, Monahan, & Slobogin, 2014; Hart & Cooke, 2013; Hart, Michie, & Cooke, 2007) have discussed the logical, methodological, and statistical barriers to defining and estimating individual-level predictions of violence risk, including predictions of violence using the PCL–R.

24. These two new and important findings concerning the validity of the PCL–R with respect to the prediction of institutional violence are likely due, at least in part, to the limited interrater reliability and substantial margin of error of PCL–R total scores.

Changes Over Time in the Evidence Base Concerning the Interrater Reliability and Predictive Validity of the PCL–R

25. Prior to the mid-2000s, the existing evidence base (i.e., body of peer-reviewed research) concerning the PCL–R was limited in important respects. There was no research supporting either the interrater reliability of the PCL–R in field settings or the predictive validity of the PCL–R with respect to serious institutional violence—that is, there was an "absence of proof" of the PCL–R's reliability and validity in these respects.

26. Since the mid-2000s, the evidence base concerning the PCL–R has expanded greatly. There is now a body of research indicating serious problems with the interrater reliability of the PCL–R in field settings and the predictive validity of the PCL–R with respect to serious institutional violence—that is, there is now "proof of absence" of the PCL–R's reliability and validity in these respects.

27. For these reasons, it is our consensus opinion that PCL–R scores cannot and should not be used to estimate the likelihood or predict that people will commit serious institutional violence. The use of PCL–R scores for such purposes is inconsistent with standards of practice in the field of forensic mental health.

(*Appendix continues*)

Attachment A

Members of the Group of Concerned Forensic Mental Health Professionals

Marcus T. Boccaccini
Department of Psychology and Philosophy, Sam Houston State University
Mark D. Cunningham
Private Practice, Seattle, Washington
David DeMatteo
Department of Psychology & Thomas R. Kline School of Law, Drexel University
Kevin S. Douglas
Department of Psychology, Simon Fraser University
Joel A. Dvoskin
Department of Psychiatry, University of Arizona College of Medicine
John F. Edens
Department of Psychological & Brain Sciences, Texas A&M University
Laura S. Guy
Department of Psychology, Simon Fraser University
Stephen D. Hart
Department of Psychology, Simon Fraser University
Kirk Heilbrun
Department of Psychology, Drexel University
Daniel C. Murrie
Institute of Law, Psychiatry, and Public Policy, University of Virginia
Randy K. Otto
Department of Mental Health Law & Policy, University of South Florida
Ira K. Packer
Department of Psychiatry, University of Massachusetts Medical School
Thomas J. Reidy
Private Practice, Monterey, California

Attachment B

Items in the Hare Psychopathy Checklist—Revised

Item
1. Glibness/superficial charm
2. Grandiose sense of self worth
3. Need for stimulation/proneness to boredom
4. Pathological lying
5. Conning/manipulative
6. Lack of remorse or guilt
7. Shallow affect
8. Callous/lack of empathy
9. Parasitic lifestyle
10. Poor behavioral controls
11. Promiscuous sexual behavior
12. Early behavioral problems
13. Lack of realistic, long-term goals
14. Impulsivity
15. Irresponsibility
16. Failure to accept responsibility for own actions
17. Many short-term marital relationships
18. Juvenile delinquency
19. Revocation of conditional release
20. Criminal versatility

Received October 25, 2019
Revision received November 25, 2019
Accepted November 26, 2019 ■

UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

UNITED STATES OF AMERICA	)	
v.	)	Criminal Action No. 01-10384-LTS
GARY LEE SAMPSON	)	

SEALED MEMORANDUM AND ORDER ON RULE 12.2 MOTIONS

September 2, 2016

SOROKIN, J.

Gary Lee Sampson pled guilty to two counts of carjacking resulting in death and was sentenced to death in 2004. The First Circuit affirmed the judgment. United States v. Sampson, 486 F.3d 13 (1st Cir. 2007). In 2011, the Court (Wolf, J.) vacated the death sentence in light of juror misconduct, and the First Circuit affirmed, ruling that Sampson is entitled to a new penalty phase trial pursuant to 28 U.S.C. § 2255. Sampson v. United States, 724 F.3d 150, 170 (1st Cir. 2013). The case was reassigned to this session of the Court on January 6, 2016.

This Order resolves eight motions in limine raising issues related to expert testimony on the subject of Sampson's mental condition, which the parties anticipate offering at trial during the defense's mitigation case and the government's rebuttal case pursuant to Federal Rule of Criminal Procedure 12.2. The motions are fully briefed and supported by voluminous exhibits as described below. Certain motions were the subject of a sealed evidentiary hearing during the week of July 25, 2016, and the Court heard oral argument on all motions at the conclusion of the hearing. The Court will address each pending motion in turn, after providing a brief summary of the relevant facts and legal standards, in the sections that follow.

TABLE OF CONTENTS

Insofar as Dr. Welner's report includes facts deemed inadmissible by prior orders of Judge Wolf or this Court, the appropriate remedy is to eliminate references to and reliance upon such facts as discussed below. Although the Court does not find them to be grounds for disqualification, the prior cases Sampson identifies in which questions have arisen surrounding Dr. Welner's compliance with court orders do provide a troubling lens through which the Court has assessed the other issues raised in Sampson's motions.[28] See Doc. No. 2345 at 16-19 (summarizing instances in which other federal trial courts, in three capital cases, have explored allegations that Dr. Welner violated orders issued pursuant to Rule 12.2).

Under these circumstances, the request for disqualification is DENIED.

3. *Psychopathy*[29]

Sampson asks the Court to preclude testimony by Dr. Welner that Sampson is a psychopath, alleging unreliability and significant prejudicial effect. Doc. No. 2345 at 20-37. In both of his 2016 reports, Dr. Welner diagnoses Sampson with psychopathy, applying the twenty criteria contained in the Hare Psychopathy Checklist-Revised ("PCL-R"), the so-called "gold standard" for making such a diagnosis. PCL-R Manual at 1; Doc. No. 2428 at 49. Dr. Welner made the same diagnosis of Sampson in 2003, and Judge Wolf precluded him from testifying about it then based on concerns about the risk of jurors considering such evidence as proof of

---

part of the government's rebuttal evidence at trial, and Dr. Welner's testimony will be circumscribed to conform with all relevant Court orders as discussed below.

[28] It is not accurate, as the government suggests, that Judge Wolf "rejected Sampson's additional arguments regarding purported 'violations' by Dr. Welner in . . . other cases." Doc. No. 2355 at 17. Rather, the transcript of the relevant hearing reveals that Judge Wolf did not view those allegations as a reason to disqualify Dr. Welner at that time, but did not rule on the merits of Sampson's assertions related to those prior incidents. Doc. No. 1873 at 83.

[29] Although Sampson discusses psychopathy and ASPD together, the record before the Court reveals material differences that warrant separate analysis of each diagnosis.

future dangerousness – a nonstatutory aggravator alleged by the government in 2003 and realleged now. United States v. Sampson, 335 F. Supp. 2d 166, 222 n.27 (D. Mass. 2004). After careful consideration, Sampson's request to bar testimony about psychopathy is ALLOWED for two reasons.

First, as this Court has said before, it generally intends to adhere to Judge Wolf's rulings from the first penalty phase trial regarding the exclusion of evidence. Doc. No. 2189 at 35-36. The government appropriately argues that the Court should not view as binding Judge Wolf's prior ruling in this regard, as the relevant evidence will be part of the government's rebuttal case, the scope of which depends on Sampson's mitigation presentation and, thus, is not the same as it was during the first trial. Doc. No. 2355 at 23. The Court agrees with this framework. However, the primary concern Judge Wolf expressed regarding evidence of psychopathy – i.e., the danger of "injecting expert comment on future dangerousness into [jurors'] discussion," Tr. of Dec. 4, 2003 Sealed Lobby Conf. at 6 – remains unchanged. Pursuant to Rule 12.2(c)(4), Dr. Welner's testimony in this case is admissible only insofar as it rebuts evidence of mental condition Sampson offers in mitigation. It may not be offered in support of the government's aggravating factors. This is true regardless how Sampson may reformulate his mental condition mitigators at his new penalty phase trial. Thus, there is no reason to believe that the circumstances of the retrial will somehow eliminate Judge Wolf's worry that jurors, even with appropriate limiting instructions, would be tempted to consider testimony that Sampson is a psychopath as evidence probative of future dangerousness, a topic upon which Dr. Welner would not be permitted to opine directly. This is especially so where evidence supporting a diagnosis of psychopathy would include specific testimony about how Sampson satisfies the PCL-R

criteria, such as "lack of remorse or guilt," "lack of empathy," and "criminal versatility." PCL-R Manual at 1.

Although Dr. Welner testified he did not use the PCL-R "as a prospective instrument" intended to predict future danger, Doc. No. 2428 at 155, the very first page of the PCL-R Manual touts the test as having an "unparalleled" and "unprecedented" ability "to predict violence." See Doc. No. 2436 at 146 (reflecting testimony by Dr. Edens about contexts in which the PCL-R is used to assess risk of reoffending). Studies have shown that jurors connect information about psychopathy and its criteria with concepts of future dangerousness.[30] Id. at 144. Not only would such a connection be impermissible within the bounds of Rule 12.2, it also would be misleading. Dr. Edens testified that the PCL-R is widely used as "a risk assessment tool" in contexts requiring decisions about whether to release certain offenders. See id. at 146 (referencing use in "civil commitment cases" and in California parole decisions). In such contexts, where decisionmakers are assessing the chance of "community recidivism," Dr. Edens said there is evidence of some relationship between high scores on the PCL-R and likelihood of rearrest following release. Id. at 144-45. But release is not an option for Sampson; the only two sentences available for the jury's consideration are death, or life in prison without the possibility of parole. See United States v. Sampson, 275 F. Supp. 2d 49, 108 (D. Mass. 2003) (Wolf, J.). According to Dr. Edens, a high PCL-R score does not meaningfully predict aggressive behavior

[30] Again, Dr. Welner's own words are revealing on this subject. He has written that "call[ing] someone a psychopath in a forensic examination" is "really like putting the mark of Cain on his or her forehead," and has described the diagnosis as "damning." Michael Welner, Hidden Diagnosis & Misleading Testimony: How Courts Get Shortchanged, 24 Pace L. Rev. 193, 203 (2003); see also Doc. No. 2345 at 31-33 (citing an article by the PCL-R checklist's creator, in which he discusses common assumptions about "psychopaths," and studies suggesting jurors are more likely to view defendants as dangerousness after hearing information about psychopathy and its criteria).

in prison – the only relevant context in which future dangerousness is at issue here. Id. Under these circumstances, Judge Wolf's reasoning remains persuasive. The Court concurs with his finding that evidence of psychopathy would fail the § 3593(c) balancing test, and further believes that the principles underlying § 2255 support applying that finding in the context of Sampson's retrial.

Second, based on the evidence offered by the parties in support of their briefs and during the evidentiary hearing on these motions, the Court concludes that testimony about psychopathy is inadmissible as rebuttal evidence under Daubert and Rule 702. Psychopathy is not a disorder listed in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition ("DSM-5"). See Doc. No. 2428 at 257 (reflecting testimony by Dr. Sanislow that "[e]very psychiatric diagnosis recognized by the field is in the DSM"); see also id. at 102-03 (noting the PCL-R creator has tried and, so far, failed to secure a place in the DSM for psychopathy). Published by the American Psychiatric Association, the DSM-5 "is a classification of mental disorders with associated criteria designed to facilitate more reliable diagnoses of these disorders." DSM-5 at xli. Psychopathy's absence from the DSM underscores its status as a "construct," as distinct from a "disorder." See Doc. No. 2436 at 6-7 (reflecting testimony by Dr. Woods that a "construct . . . hasn't really gained the number of consistent symptoms" to be considered a syndrome or disorder).

Despite its increased popularity among psychiatrists over the past decade or so,[31] the record before the Court demonstrates that substantial questions exist as to the reliability and validity of psychopathy as a diagnosis. Dr. Edens – a Professor of Psychology at Texas A&M

[31] To the extent the test has gained popularity due to growing use as a tool to assess likelihood of recidivism as part of parole decisions and in other similar contexts, such increased popularity does not establish reliability or general acceptance for present purposes.

University whose research and writing largely focus on psychopathy and related disorders, see Curriculum Vitae of John F. Edens, Ph.D., Edens Binder ("Edens CV") – described his own studies and the evolution of his thinking on the subject, citing problems with inter-rater reliability (i.e., the chance that two separate doctors will score the PCL-R the same when assessing the same examinee) and a ***large*** standard error measurement (i.e., the range around a single rater's score within which an examinee's true score falls). Doc. No. 2436 at 95-112. Other witnesses testified about similar concerns with the diagnosis. E.g., id. at 26-29 (reflecting testimony by Dr. Woods about concerns including lack of exclusion criteria, confirmatory bias, and lack of guidance defining criteria). The Court credits the defense experts' testimony in this regard.

Dr. Welner was the only testifying expert to endorse use of the PCL-R without reservation. His testimony, however, did not answer the concerns identified by the defense experts, at least one of whom has studied the test over two decades and published numerous peer-reviewed articles about its strengths and limitations. See Edens CV at 3-20. Furthermore, his application of the PCL-R in this case did not incorporate at least two techniques recommended in the test's manual for avoiding biases and enhancing reliability. See PCL-R Manual at 18, 20 (recommending "strongly" the "averaging [of] PCL-R scores of two or more independent raters" in order to increase test reliability and reduce measurement error,[32] and advising raters to score each item separately and "make written notes on a separate sheet of paper to justify [each] rating"[33]). That other courts have permitted experts (including Dr. Welner) to

[32] Here, not only was Dr. Welner the ***only*** rater, he also scored Sampson using the PCL-R three separate times. See Doc. No. 2322-1 at 223-24, 293. This practice, it seems to the Court, would tend to magnify any biases and decrease reliability, rather than protecting against such problems.

[33] The PCL-R Manual acknowledges "two common rating biases" – a "halo effect (basing each item score on a global impression of the individual, perhaps unduly influenced by the nature of

opine on the subject of psychopathy in other cases does not overcome this Court's concerns about the reliability of such testimony in this case. See Doc. No. 2355 at 20-21 (citing cases in which courts permitted psychopathy evidence).[34]

In sum, psychopathy evidence – including testimony by Dr. Welner as to his administration and scoring of the PCL-R and his diagnosis of Sampson as a psychopath – will not be admitted at Sampson's retrial. The Court precludes such evidence (a) under the § 3593(c) balancing test due to an overwhelming risk of prejudice and of misleading the jury as to the purpose of the evidence; (b) in its § 2255 discretion, based on Judge Wolf's decision to exclude such evidence in Sampson's first trial for reasons that apply with equal force now; and (c) as unreliable under Daubert and Rule 702, based primarily on the concerns identified by Dr. Edens in his testimony and expert declaration.[35]

4. *Antisocial Personality Disorder*

Sampson also seeks to preclude testimony by Dr. Welner that Sampson suffers from ASPD. Doc. No. 2345 at 20-37. He makes the same claims of unreliability and prejudice as to

---

the offenses)," and a "'nice- or bad-guy' bias (rating all items low or high)." As far as the Court is aware, no contemporaneous notes exist showing Dr. Welner's justifications for his scoring. In fact, although his reports suggest he scored Sampson three separate times, it appears he completed the scoring form that is to be used when administering the test only once – on July 20, 2016, the date of his revised report, and a year after his most recent examination of Sampson. See Doc. No. 2355-5.

[34] In at least one case cited by the government, psychopathy evidence was treated as "probative of . . . future dangerousness." United States v. Lee, 274 F.3d 485, 495 (8th Cir. 2001).

[35] The Court notes – but does not rely upon as a basis for this decision – a discrepancy in the record as to whether Dr. Welner has received training in administering the PCL-R. Compare Doc. No. 2428 at 49, 143-44 (reflecting testimony before this Court that Dr. Welner is "trained specifically on the PCL-R" and took "the training program . . . tout[ed by the test's creator] as being necessary in order to properly administer th[e] test" in 2000 or 2001), with Doc. No. 2383-1 at 3, 6 (reflecting testimony from another case in which Dr. Welner denied having received training in scoring the PCL-R); see also Doc. No. 2383 (providing Dr. Welner's explanation for the differing testimony).

from Dr. Welner's testimony at Sampson's first trial). However, where the allegedly independent "peer review" process concededly was not used here, any description of that process (even in describing Dr. Welner's qualifications) would be irrelevant and misleading in this case. Accordingly, to the extent the "Peer Review" Motion (Doc. No. 2336) seeks to preclude testimony characterizing the Forensic Panel generally or its members specifically as "independent" and/or as utilizing "internal oversight and peer review," the motion is ALLOWED.

IV. CONCLUSION

For the foregoing reasons, the Court ORDERS as follows:

1) the Gur Motion (Doc. No. 2325) is DENIED;
2) the Aguirre Motion (Doc. No. 2331) is DENIED;
3) the Malingering Motion (Doc. No. 2339) is ALLOWED IN PART and DENIED IN PART as described above;
4) the Welner Motion (Doc. No. 2333) is ALLOWED IN PART and DENIED IN PART as described above, and on or before September 23, 2016, the government shall provide to the defense and the Court a revised expert report prepared by Dr. Welner in conformance with the rulings set forth in Discussion § III(D) above;
5) the Parameters Motion (Doc. No. 2322) is DENIED;
6) the Judicial Determinations Motion (Doc. No. 2328) is DENIED;
7) The Expert Designation Motion (Doc. No. 2334) is DENIED; and
8) the "Peer Review" Motion (Doc. No. 2336) is ALLOWED IN PART and DENIED IN PART.

The motion papers and supporting documents as to each of these motions are sealed, as is this Memorandum and Order.[59] These items will remain sealed until the conclusion of the trial for two reasons. First, all of the relevant motions arise from evidence that is within the scope of Rule 12.2, which is aimed at protecting a defendant's Fifth Amendment rights. If Sampson elects not to offer evidence of mental condition at trial, none of the issues raised in these motions or the information underlying them will be admitted at trial. His Fifth Amendment protections in this regard are only completely waived when he offers his own expert mental condition evidence at trial. Until then, sealing is required consistent with Rule 12.2. Second, the motions, the supporting documents, and this Memorandum and Order contain certain information that will not be admissible during the trial in this matter, as well as other information the admissibility of which will depend on how the defense elects to shape its mitigation presentation at trial. As such, continued sealing is appropriate. See In re Globe Newspaper Co., 729 F.2d 47, 55 (1st Cir. 1984). If mental condition evidence is offered at trial, the Court anticipates unsealing all motions, briefs, and exhibits related to these motions, as well as this Memorandum and Order, after the conclusion of the trial.

SO ORDERED.

/s/ Leo T. Sorokin
Leo T. Sorokin
United States District Judge

[59] The parties may share copies of this Memorandum and Order with the relevant expert witnesses to assist in preparations for trial. Those experts who receive or review this Memorandum and Order may not disseminate it and are bound by the sealing order.

APPENDIX A

Expert Reports & Declarations

1. Michael Welner, October 31, 2003 Report (Doc. No. 2333-2)
2. Michael Welner, April 4, 2016 Report (Doc. No. 2322-1, pp. 188-277)
3. Michael Welner, June 20, 2016 Report (Doc. No. 2322-1, pp. 279-333)
4. Michael Welner, July 13, 2016 Declaration, (Doc. No. 2355-1)
5. Thomas Guilmette, April 3, 2016 Report (Doc. No. 2322-1, pp. 29-140)
6. Thomas Guilmette, June 20, 2016 Report (Doc. No. 2322-1, pp. 142-186)
7. Thomas Guilmette, July 13, 2016 Declaration (Doc. No. 2355-3)
8. Geoffrey Aguirre, August 4, 2015 Report (Doc. No. 2322-1, pp. 3-16)
9. Geoffrey Aguirre, June 20, 2016 Report (Doc. No. 2322-1, pp. 18-27)
10. Erin Bigler, April 6, 2016 Report (Doc. No. 2322-2, pp. 8-9)
11. Erin Bigler, June 20, 2016 Report (Doc. No. 2322-2, pp. 11-17)
12. Erin Bigler, July 22, 2016 Letter (Doc. No. 2369-1)
13. John Edens, June 17, 2016 Declaration (Doc. No. 2322-2, pp. 19-35)
14. James Gilligan, March 22, 2010 Declaration (Doc. No. 1041-194, pp. 2-30)
15. James Gilligan, April 8, 2016 Report (Doc. No. 2322-2, pp. 37-49)
16. Ruben Gur, May 8, 2009 Report (Doc. No. 2325-2)
17. Ruben Gur, November 11, 2009 Report (Doc. No. 2354-3, pp. 26-28)
18. Ruben Gur, § 2255 Declaration (Doc. No. 2325-1)
19. Ruben Gur, March 5, 2010 Supplemental Declaration (Doc. No. 2354-4)
20. Ruben Gur, April 7, 2016 Report (Doc. No. 2322-2, pp. 51-55)
21. Ruben Gur, May 17, 2016 Letter (Doc. No. 2322-2, pp. 57-82; Doc. No. 2322-3, pp.1-5)
22. Ruben Gur, June 6, 2016 Letter (Doc. No. 2322-3, pp. 14-16)
23. Ruben Gur, June 19, 2016 Letter (Doc. No. 2322-3, pp. 6-12)
24. Ruben Gur, July 15, 2016 Letter (Doc. No. 2356-9)
25. Ruben Gur, July 15, 2016 Response (Doc. No. 2356-10)
26. Leslie Lebowitz, March 28, 2010 Declaration (Doc. Nos. 1041-201, -202)
27. Leslie Lebowitz, April 8, 2016 Report (Doc. No. 2322-3, pp. 18-19)
28. Leslie Lebowitz, June 20, 2016 Report (Doc. No. 2322-3, pp. 21-30)
29. Michael Lipton, July 20, 2015 Report (Doc. No. 2322-3, pp. 32-34)
30. Michael Lipton, May 30, 2016 Report (Doc. No. 2322-3, pp. 36-39)
31. James Merikangas, April 8, 2016 Report (Doc. No. 2322-4, pp. 1-3)
32. Paul Moberg, April 7, 2016 Report (Doc. No. 2322-4, pp. 5-16)
33. Paul Moberg, June 20, 2016 Report (Doc. No. 2322-4, pp. 18-22)
34. Charles Sanislow, June 20, 2016 Report (Doc. No. 2322-4, pp. 30-59)
35. George Woods, March 24, 2010 Declaration (Doc. No. 1041-203, -204, -205)
36. George Woods, April 8, 2016 Report (Doc. No. 2322-4, pp. 61-101)
37. George Woods, June 20, 2016 Report (Doc. No. 2322-4, pp. 103-133)

Other Exhibits to Briefs

38. Ruben C. Gur et al., "Behavioral Imaging" – A Procedure for Analysis and Display of Neuropsychological Test Scores: I. Construction of Algorithm and Initial Clinical Evaluation, Neuropsychiatry, Neuropsychology, & Behavior Neurology, Vol. 1, No. 1, p. 53 (1988) (Doc. No. 2325-4)

39. Ruben C. Gur et al., "Behavioral Imaging" – II. Application of the Quantitative Alogrithm to Hypothesis Testing in a Population of Hemiparkinsonian Patients, Neuropsychiatry, Neuropsychology, & Behavior Neurology, Vol. 1, No. 2, p. 87 (1988) (Doc. No. 2325-5)
40. Ruben C. Gur et al., "Behavioral Imaging" – III. Interexpert Agreement and Reliability of Weightings, Neuropsychiatry, Neuropsychology, & Behavior Neurology, Vol. 3, No. 2, p. 113 (1990) (Doc. No. 2325-6)
41. Trial Transcript, Vol. 43, United States v. McCluskey, No. 10-cr-2734-JCH (D.N.M. Oct. 24, 2013) (Doc. No. 2325-7)
42. Hearing Transcript, United States v. Northington, No. 07-cr-550-RBS (E.D. Pa. Oct. 17, 2012) (Doc. No. 2325-8)
43. Ronald A. Yeo et al., Neuropsychological Methods of Localizing Brain Dysfunction: Clinical Versus Empirical Approaches, Neuropsychiatry, Neuropsychology, & Behavior Neurology, Vol. 3, No. 4, p. 290 (1990) (Doc. No. 2325-9)
44. William J. Lynch, Letter to Assistant District Attorney Martin Murray (Jan. 6, 2005) (Doc. No. 2325-11)
45. Deposition Transcript, United States v. Duong, No. 01-cr-20154-JF (N.D. Cal. Dec. 3, 2010) (Doc. No. 2325-13)
46. Partial Jury Trial Transcript, United States v. Green, No. 06-cr-19-TBR (W.D. Ky. May 19, 2009) (Doc. No. 2325-14)
47. Partial Hearing Transcript, United States v. Montgomery, No. 05-cr-6002-GAF (W.D. Mo. Sept. 4-5, 2007) (Doc. Nos. 2325-15, -16)
48. Partial Trial Transcripts, Commonwealth v. Chism, Nos. 2013-1446, 2013-1447, 2014-0109 (Mass. Super. Ct. Essex Cty. Dec. 3 & 10, 2015) (Doc. No. 2325-17)
49. Ruben C. Gur Curriculum Vitae (Doc. No. 2356-1)
50. Letters of Reference re: Gur (Doc. No. 2356-2)
51. Ruben C. Gur et al., Response to Yeo et al.'s Critique of Behavioral Imaging, Neuropsychiatry, Neuropsychology, & Behavior Neurology, Vol. 3, No. 4, p. 304 (1990) (Doc. No. 2356-6)
52. Exhibit 63-65 in Support of Petition for Executive Clemency – Donald Jay Beardslee (Doc. No. 2356-7)
53. David R. Roalf Letter, July 8, 2016 (Doc. No. 2356-8)
54. Geoffrey K. Aguirre Curriculum Vitae (Doc. No. 2354-5)
55. Medical Records re 2003 MRI of Gary Sampson (Doc. No. 2354-6)
56. Medical Records re 2009 MRI of Gary Sampson (Doc. No. 2354-7)
57. Transcript of Interview of Gary Sampson by Michael Welner, October 29, 2003 (Doc. No. 2333-1)
58. Cory S. Flashner Letter to William E. McDaniels (June 25, 2015) (Doc. No. 2333-4)
59. Transcript of Interview of Gary Sampson by Michael Welner, July 16, 2015 (Doc. No. 2333-5)
60. Transcript of Interview of Gary Sampson by Michael Welner, July 17, 2015 (Doc. No. 2333-6)
61. E-mails between Zachary Hafer and Michael Burt (May 2016) (Doc. No. 2333-8)
62. Leigh D. Hagan and Thomas J. Guilmette, DSM-5: Challenging Diagnostic Testimony, International Journal of Law & Psychiatry 42-43 (2015) 128-134 (Doc. No. 2333-13)

63. Sanford L. Drob et al., Clinical and Conceptual Problems in the Attribution of Malingering in Forensic Evaluations, J. Am. Acad. Psychiatry Law 37:98-106 (2009) (Doc. No. 2341-7)
64. Michael F. Martelli et al., Masquerades of Brain Injury Part III: Critical Evaluation of Symptom Validity Testing and Diagnostic Realities in Assessment, Journal of Controversial Medical Claims, Vol. 9 No. 2 p.19 (2002) (Doc. No. 2341-8)
65. Richard Rogers et al., Assessment of Malingering with Repeat Forensic Evaluations: Patient Variability and Possible Misclassification on the SIRS and Other Feigning Measures, J. Am. Acad. Psychiatry Law 38:109-14 (2010) (Doc. No. 2341-9)
66. Michael Welner Curriculum Vitae (Doc. No. 2355-2)
67. Thomas J. Guilmette Curriculum Vitae (Doc. No. 2355-4)
68. Hare PCL-R: 2nd Edition Score Sheets of Gary Sampson by Michael Welner (June 20, 2016) (Doc. No. 2355-5)
69. Jennifer Cox et al., The Effect of the Psychopathy Checklist – Revised in Capital Cases: Mock Jurors' Responses to the Label of Psychopathy, Behav. Sci. Law 28: 878-891 (2010) (Doc. No. 2355-6)
70. John F. Edens Curriculum Vitae (Doc. No. 2355-7)
71. Mark D. Cunningham and Thomas J. Reidy, Antisocial Personality Disorder Versus Psychopathy as Diagnostic Tools (1998) (Doc. Non. 2355-8)

Evidentiary Hearing Exhibits

72. Thomas J. Guilmette 12.2 Witness Binder ("Guilmette Binder")
73. Dr. Michael Welner 12.2 Witness Binder ("Welner Binder")
74. Charles A. Sanislow 12.2 Witness Binder ("Sanislow Binder")
75. Dr. George Woods 12.2 Witness Binder ("Woods Binder")
76. Dr. John F. Edens 12.2 Witness Binder ("Edens Binder")
77. Hare Psychopathy Checklist – Revised (PCL-R): 2nd Edition, Technical Manual ("PCL-R Manual")
78. David DeMatteo et al., Investigating the Role of the Psychopathy Checklist-Revised in United States Case Law, 20 Psychol., Pub. Pol'y & L. 96 (2014)
79. M.E. Thomas, Confessions of a Sociopath: A Life Spent Hiding in Plain Sight (2014) (excerpts)
80. John F. Edens et al., Bold, Smart, Dangerous & Evil: Perceived Correlates of Core Psychopathic Traits Among Jury Panel Members, 7 Personality & Mental Health 143 (2013)
81. John F. Edens et al., Use of the Personality Assessment Inventory to Assess Psychopathy in Offender Populations, 12 Psychol. Assessment 132 (2000)
82. Darrel B. Turner et al., Jurors Report that Risk Measure Scores Matter in Sexually Violent Predator Trials, but that Other Factors Matter More, 33 Behav. Sci. & L. 56 (2015)
83. Jennifer Cox et al., The Effect of the Psychopathy Checklist – Revised in Capital Cases: Mock Jurors' Responses to the Label of Psychopathy, 28 Behav. Sci. & L. 878 (2010)
84. Welner Excluded Evidence Tracking Chart (Doc. No. 2390)
85. Zachary Hafer Post-Hearing Letter and attachments, Aug. 1, 2016 (Doc. No. 2383)
86. Michael Burt Post-Hearing Letter, Aug. 2, 2016 (Doc. No. 2387)

FORT WORTH INDEPENDENT SCHOOL DISTRICT
PSYCHOLOGICAL SERVICES DEPARTMENT
3210 WEST LANCASTER
FORT WORTH, TEXAS 76107

INTELLECTUAL SUMMARY

Name: Dustin Jones School: V Williams
DOB: 2-15-79 Grade: 2nd
Age: 7-9

Date of Evaluation: 5-7-87

Reason for Referral: ✓ Special Education, _____ Parent, _____ Other.

Previous Test Results: _____

Tests and Results:

WISC-R: Verbal IQ 101, Performance IQ 98, Full Scale IQ 100.

Verbal Subtests	Scaled Scores	Performance Subtests	Scaled Scores
Information	8	Picture Completion	9
Similarities	13	Picture Arrangement	10
Arithmetic	9	Block Design	8
Vocabulary	9	Object Assembly	9
Comprehension	12	Coding	13
Digit Span	13		

Slosson: Stanford-Binet: IQ _____, MA _____.

CMMS: Raw Score _____, Age Deviation _____, Stanine _____, Maturity Index _____, Percentile Rank _____.

WRAT-R:	Raw Score	Standard Score	Percentile	Grade Equivalent
Reading				
Spelling				
Arithmetic				

Woodcock-Johnson:	Grade Level	Percentile Rank-Grade/Age	Standard Score
Reading	1.8	9 grade	80
Math	2.2	20 grade	87
Written Language	2.0	20	87
Knowledge			

Bender Visual-Motor Gestalt Test: _____ Emotional indicators _____

Human Figure Drawing Test: _____ Emotional indicators _____

Other test results: _____

Interpretations and Recommendations: Based on testing with the WISC-R Dustin's intellectual functioning is in the average range. His areas of strength appear to be in concrete

APPLIED NEUROPSYCHOLOGY, *16*: 91–97, 2009

ISSN: 0908-4282 print/1532-4826 online
DOI: 10.1080/09084280902864329

Psychology Press
Taylor & Francis Group

Interpretation of Intelligence Test Scores in *Atkins* Cases: Conceptual and Psychometric Issues

Frank M. Gresham
Department of Psychology, Louisiana State University, Baton Rouge, Louisiana

So-called *Atkins* cases refer to individuals who have been sentenced to death for capital crimes who claim that the death penalty constitutes "cruel and unusual punishment" under the Eighth Amendment. Psychological testimony is influential because this testimony strikes at the very core issue in these cases; namely, whether or not the individual is mentally retarded. Despite the importance of psychological testimony, courts have not been made to understand the subtleties and complexities of the issues in diagnosing mental retardation. Five such issues are discussed in this article: (a) the nature of intellectual functioning, (b) the Flynn Effect, (c) measurement error, (d) practice effects, and (e) the nature of school "diagnoses." Examples of each of these issues are illustrated with an actual Atkins case (*Walker v. True*, 2006).

Key words: *Flynn Effect, intelligence, measurement error, psychometric*

Around midnight on August 16, 1996, Daryl Renard Atkins and an accomplice (William Jones) abducted Erich Nesbitt with a semiautomatic handgun and robbed him of his money. Subsequently, they drove Nesbitt to an ATM and forced him to withdraw cash. He was then taken to an isolated spot where he was shot eight times and killed. During trial, both Atkins and Jones testified and confirmed each other's account of the incident, except that Jones' testimony was considered more credible than Atkins'. In fact, Atkins' court testimony was substantially inconsistent with the testimony he gave police upon his arrest, whereas Jones declined to make a statement to authorities upon his arrest (Miranda Rights). During the penalty phase of the trial, the defense relied on Dr. Evan Nelson, a forensic psychologist, who had evaluated Atkins prior to trial and concluded that he was mildly mentally retarded based on a review of school and court records and a tested full scale IQ of 59 on the Wechsler Adult Intelligence Scale-III (WAIS-III). Atkins, however, was sentenced to death, and Jones plea bargained with the prosecution in return for testimony against Atkins and was spared the death penalty.

At a second sentencing hearing, another forensic psychologist, Dr. Stanton Samenow, expressed the opinion that Atkins was not mentally retarded and was functioning in the range of "average" intelligence. This opinion was based on two interviews with Atkins, a review of school records, the Wechsler Memory Scale (Wechsler, 1972), and interviews with correctional officers. Dr. Samenow did not administer an intelligence test but opined that Atkins' poor academic performance while in school was due to his frequent inattention and his overall tendency toward noncompliance in school.

How can two board-certified, licensed, forensic psychologists come to two diametrically opposed opinions regarding the presence or absence of mental retardation? Atkins' measured intelligence was over 2.7 standard deviations below the mean, which almost pushed him into the moderate range of mental retardation (American Psychiatric Association, 2000). Despite this fact, the prosecution's psychologist considered Atkins to be of average intelligence. This finding, as is demonstrated throughout this special issue, is neither unusual nor unexpected for a variety of reasons that will be discussed

Address correspondence to Frank M. Gresham, Department of Psychology, Louisiana State University, Baton Rouge, LA 70803. E-mail: gresham@lsu.edu

in this article. I start with a very brief overview of mental retardation, particularly mild mental retardation, and continue with a discussion of interpretative and psychometric issues in the assessment of intelligence. The article concludes with recommendations for psychologists who may one day find themselves as experts in *Atkins* cases.

MENTAL RETARDATION

Mental retardation is defined by most organizations and states as significantly subaverage intellectual functioning that concurrently exists with deficits in adaptive behavior and which has an onset prior to age 18. The *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 2000) specifies that significantly subaverage intellectual functioning should be two standard deviations below the mean, however, it acknowledges that the existence of five points in measurement error should be considered in making a diagnosis of mental retardation. As such, it is possible to diagnose an individual as having mental retardation with an IQ up to 75 if they also have substantial deficits in adaptive behavior. Adaptive behavior refers to how well an individual copes with life demands and how well they meet the standards of personal independence expected of someone in their age group, sociocultural background, and community setting (APA, 2000). DSM-IV specifies four degrees of severity for mental retardation: mild mental retardation (IQ 50–55 to 70–75), moderate mental retardation (IQ 35–40 to 50–55), severe mental retardation (IQ 20–25 to 35–40) and profound mental retardation (IQ below 20 or 25). As will be described later, the debate in the *Atkins* cases has never been about individuals with moderate, severe, or profound mental retardation. It has always been about persons who might be considered to have *mild mental retardation*.

The American Association on Intellectual and Developmental Disabilities (AAIDD, 2002) defines an intellectual disability as being characterized by significant limitations both in intellectual functioning and in adaptive behavior as expressed in conceptual, social, and practical adaptive skills and originates before the age of 18. Similar to DSM-IV, significant limitations in intellectual functioning is defined as performance that is two standard deviations below the mean (70–75 and below); imitations in adaptive behavior is defined as performance that is at least two standard deviations below the mean in one of the three adaptive behavior domains (conceptual, social, or practical) or a total adaptive behavior (composite) score on a standardized adaptive behavior measure (see Greenspan and Reschly's discussion of adaptive behavior, this issue). Unlike DSM-IV, however, AAIDD does not classify mental retardation by severity (mild, moderate, severe, or profound), but rather uses the concept of levels of supports needed to promote the development, education, interests, and personal well-being of an individual with intellectual disability.

An extremely important issue in *Atkins* cases that is often misunderstood by the courts is the nature of mild mental retardation (MMR) as being distinct from more severe forms. First, MMR has no identified or specified biological etiology, whereas more severe forms of mental retardation often have an identified biological etiology (e.g., Down syndrome, Fragile X syndrome, and microcephaly). Second, MMR is most often diagnosed only at school entry or shortly thereafter, whereas severe forms of mental retardation are often diagnosed at birth or shortly thereafter. Third, adaptive behavior functions of persons with MMR may be adequate in some areas (e.g., practical skills), but severely deficient in others (e.g., conceptual). Individuals with severe forms of mental retardation almost always have pervasive adaptive behavior deficits. Finally, persons with MMR may "blend" into society after school exit (Edgerton, 1993) and appear to function normally in community settings, whereas persons with severe forms of mental retardation will always "stand out" because of their physical anomalies and severely pervasive intellectual and adaptive behavior deficits. It is apparent that the courts have a preconceived notion of what mental retardation looks like that is inconsistent with what MMR looks like to professionals in the field who have training and experience in the field of mental retardation. Unfortunately, this bias is often perpetuated by forensic experts who testify for the prosecution, who, more often than not, have little or no training in the field of mental retardation.

INTERPRETIVE ISSUES IN INTELLECTUAL ASSESSMENT

The remainder of this article will discuss various interpretive issues in intellectual assessment that courts have failed to understand or consider in deciding *Atkins* cases. These interpretive issues are: (a) the nature of intellectual functioning, (b) the Flynn effect, (c) the concept of measurement error, (d) practice effects, and (e) the effect of school diagnoses. Each of these issues will be illustrated with actual *Atkins* cases and court decisions.

Nature of Intellectual Functioning

A major issue confronting the courts in *Atkins* cases resides in their understanding (or misunderstanding) of what intelligence tests measure and how well they

measure the construct of intelligence. The courts have a difficult time comprehending that in a psychometric world; an individual can have more than one true score. For example, suppose an individual is administered a WAIS-III, a Stanford Binet Intelligence Scale-IV (SB-V), and a Woodcock-Johnson Test of Cognitive Abilities-III (WJ Cognitive-III). All three tests yield an overall or composite intelligence score, and an individual taking all three tests will have three true scores, one for each test.

In classical test theory, an individual's true score on any attribute is entirely dependent on the measurement process that is used. In the biological and physical sciences, an individual can have only one true score and that score is independent of the measurement process that is used. This is known as the absolute true score (Crocker & Algina, 1986). For example, a laboratory may analyze an individual's DNA as part of evidence presented in court in a capital case. Individuals have only one true score for their DNA, and the courts have come to understand this phenomenon. However, different labs may obtain different results in their DNA analyses and thus errors of measurement occur. This does not alter the fact that only one true score exists, and different labs would never average the results of various lab tests to derive a true score. Yet, this is precisely how we interpret true scores on psychological measures of intelligence and other attributes.

An *Atkins* case in which I testified brings this interpretive difficulty to light. Darick DeMorris Walker was convicted of two capital murders and sentenced to death in Virginia. Walker claimed that the death penalty violated his Eight Amendment rights that protect him from "cruel and unusual punishment" because he is mentally retarded. Walker had a history of below-average intelligence and a school history of being placed into special education classrooms. Eventually, Walker dropped out of school in the eighth grade with substantial deficits in reading and math skills and a long school history of disruptive/noncompliant behavior.

Throughout his life, Walker has been administered no less than seven intelligence tests, each producing different results. What is particularly notable in these results is the disparity between Walker's crystallized and fluid intelligence. On the various Wechsler tests, Walker's Verbal IQ ranged from 70 to 87 with a median of 78. On various measures of fluid intelligence, his scores ranged from 61 to 68 with a median IQ score of 63. The question before the court was whether or not these scores were indicative of mental retardation. There are two answers to this question which, as expected, confused rather than enlightened the court. If one takes the crystallized measures as being indicative of mental retardation, it is clear that Walker is not mentally retarded. If one takes the fluid measures as indicators of mental retardation, Walker is, clearly, mildly mentally retarded.

One approach that could be taken would be to argue that different measures of intelligence have different *g loadings*, or that they vary in how well they measure a general intelligence factor. It is well established that measures of crystallized intelligence (vocabulary, verbal abstract reasoning, and general information) have much higher *g loadings* than most measures of fluid intelligence. As such, it could be argued that measures of crystallized intelligence in most circumstances provide better estimates of *g* than most measures of fluid intelligence. This, however, could be disputed on the basis that some measures of fluid intelligence have *g loadings* approaching loadings that are produced by measures of crystallized intelligence (Keith, 2005).

Apart from this argument, the U.S. District Court (Eastern District) ruled against Walker, stating that he failed to show by a preponderance of the evidence that he is mentally retarded. His case was appealed to the U.S. Fourth Circuit Court of Appeals which vacated and remanded the District Court's judgment and granted Walker an evidentiary hearing to determine whether he is mentally retarded under Virginia law. It further ordered that the district court should consider all relevant evidence pertaining to the developmental origin, intellectual functioning, and adaptive behavior aspects of Walker's claim.

Flynn Effect

It is well established that there has been a substantial increase in measured intelligence test performance over time because IQ test norms become obsolete. As such, intelligence test norms have to periodically be recalibrated to maintain their accuracy in reflecting an individual's level of intelligence. The general upward trend in IQ scores has become known as the *Flynn Effect*, named after James Flynn who first documented this phenomenon (Flynn, 1984). Based on his extensive review of the literature, Flynn established that Americans gain approximately 0.3 IQ points per year or 3 points per decade in measured intelligence. Thus, an IQ test normed in 1972 would reflect a 10.8 point gain in IQ today ($36 \times 0.3 = 10.8$ points).

The Flynn Effect has a substantial influence on the number of persons who might be classified as mentally retarded using a specified cutoff score (Ceci, Scullin, & Kanaya, 2003). For example, if you used the WISC-R that was normed in 1972 and specified a cutoff score of 70 and below, you would identify 2.27% of the population as being mentally retarded using the intellectual criterion. However, if you used the WISC-III that was normed in 1989, you would identify 5.48% of the population as being mentally retarded—more than double the

prevalence rate based on a normal distribution. Based on the Flynn Effect it is not unusual for an individual's IQ score to fluctuate above and below a specified IQ cutoff that most states used to determine eligibility for the death penalty (Kanaya, Ceci, & Scullin, 2003).

Flynn (2006) has argued that an individual's true IQ score does not change over time, only the norms change. For instance, suppose you test a girl at age eight with the WISC-R and she obtains an IQ score of 74. You retest that same girl at age 12 with the WISC-III and she obtains an IQ score of 69. There is a five-point difference between these two IQ scores, with one score being above the level for mental retardation and the other score being below that level. The girl's intelligence, however, did not change, only the norms changed, separated by 17 years.

The Flynn Effect differentially affects certain Wechsler scores. For instance, the effect is rather large for Similarities and Block Design and nonexistent for Vocabulary and Information (Flynn, 2006). One could argue that Similarities and Block Design have rather high *g* loadings (.81 and .70, respectively), therefore this must reflect "real" changes in general intelligence. However, the two subtests that are considered the best single measures of *g* (Vocabulary and Information) remain unchanged by the Flynn Effect.

In summary, Flynn argues that intelligence has not changed over time, and that changes in measured IQ reflect the fact that norms start becoming obsolete the day they are collected. If this is true, then it could be argued that the Flynn Effect is irrelevant in determining an individual's eligibility for the death penalty because it does not address the level of intelligence, but rather the accuracy of norms that are not a part of any definition of mental retardation. However, states use IQ *scores* which are inextricably and directly dependant on norms for their meaning. These scores often are rigidly adhered to by many states (e.g., Virginia) to determine a person's eligibility for the death penalty. The view that the Flynn Effect does not reflect real changes in intelligence is moot because the courts often use an absolute level of intelligence ($IQ < 70$) to determine whether an individual is eligible for capital punishment.

Measurement Error

It is obvious to any well-trained psychologist that all measurement contains error, but this is far from obvious to the courts in deciding *Atkins* cases. For example, in *Walker v. True* (2006) the United States District Court stated that use of the standard error of measurement (SEM) to lower an IQ score could just as likely be used to raise an IQ score, and that the use of such as statistic is inherently "speculative." Clearly, there is nothing speculative in the standard error of measurement given that it is entirely dependent on the reliability of the test that is used to obtain a score. The concept of measurement error goes back to the notion of a psychometric true score versus an absolute true score described earlier—a concept that courts have a difficult time understanding. Experts for the defense in *Atkins* cases have been unsuccessful in making courts understand the band of error concept (plus or minus the SEM) and the notion of a psychometric true score that falls within this band of error. Experts for the prosecution have often downplayed the importance of measurement error in these cases because it diminishes the credibility of their testimony (*Walker v. True,* 2006).

An issue relating to measurement error in these cases is the selection of the most appropriate estimate of measurement error: should it be based on internal consistency estimates, stability estimates, or both? Internal consistency estimates will almost always yield higher reliability estimates and thus will produce lower SEMs than stability estimates because stability coefficients are almost always lower.

These two estimates of measurement error reflect two different interpretations of test scores. An internal consistency estimate is based on the average interitem correlation in a test and reflects the ratio of true score variance to total variance (i.e., the reliability index), and the square root of this index is the reliability coefficient (Suen, 1990). As such, this statistic reflects how much error is contained in the obtained score and how well that score estimates the true score. This is known as the *coefficient of internal consistency*. Errors of measurement based on stability estimates reflect the fluctuations in test scores obtained at two points in time.

The problem in classical test theory is that one can have more than one reliability coefficient and thus have more than one standard error of measurement. This is inherently self-contradictory (Suen, 1990) and therefore is more likely to confuse than inform the courts. Conceptually, what is needed is a *coefficient of precision* (Coombs, 1950), which is defined as the correlation between test scores when examinees respond to the same test items (internal consistency) over time (stability) and there are no changes in examinees over time. Unfortunately, this coefficient is a theoretical entity in classical test theory and no completely defensible way of calculating it is possible. Perhaps the best that can be done at this time is to indicate that SEMs based on internal consistency estimates contain an individual's true score at one point in time, whereas SEMs based on stability estimates contain an individual's true score over repeated testings.

Practice Effects

In *Atkins* cases it is likely that defendants have been administered intelligence tests repeatedly; often

beginning in their school years. This was true in *Atkins, Walker v. True*, and *Green v. Johnson*. School records in all of these cases show that these defendants began taking intelligence tests relatively early in their school careers because they were referred to special education. *Walker* had taken seven intelligence tests by the time his case came before the United States Eighth District Court. One argument in *Walker v. True* made by the defense was that his IQ scores should be adjusted downward, in part, because of well-known *practice effects* due to repeated administrations of the same test. The Court ruled, however, in *Walker v. True* that "Petitioner has failed to present evidence that such an adjustment would be anything other than speculation" (p. 8).

Practice effects refer to gains in test scores on intelligence tests that occur when an individual is retested on the same or similar instruments. This is not a speculation but rather a well-established empirical fact. These gains are due to having been exposed to the same or very similar test items and not due to any specific performance feedback given by examiners. Practice effects for the various Wechsler scales from ages 5 to 50 years show median gains in Verbal IQ of 3 points, Performance IQ of 9 points, and Full Scale IQ of 7 points (Kaufman, 2003). Walker had taken the WISC-R three times before the age of 18 and the WAIS-III twice after the age of 18. Thus, the practice effects on Wechsler scales beginning at age 9 to 20 (his last WAIS-III) must have been quite substantial, thereby producing inflated IQ scores.

In *Atkins* cases the courts must be made to understand the average practice effect gains in IQ scores and how these artificially inflated test scores produce an overestimate of an individual's true score. This is particularly true when experts from either side administer the same test within relatively short periods of time, because the shorter the retest interval, the larger the practice effect. If we apply the median practice effect to Walker's median Full Scale IQ, his IQ goes from 76 to 69; not considering measurement error. Quite clearly, this is extremely important in *Atkins* cases, particularly in states that inflexibly adhere to $IQ < 70$ standard for mental retardation.

SCHOOL DIAGNOSES

A major source of evidence used by courts in *Atkins* cases is the documentation of whether or not the defendant had ever been identified by a school as mentally retarded. This is considered an essential piece of evidence, given that one of the eligibility prongs for a diagnosis of mental retardation is onset prior to age 18. In *Walker v. Virginia*, the defendant received special education services during elementary school, first under the label of "learning disability" and later under the label of "emotionally disturbed." Walker was never labeled as being mentally retarded by the Richmond, Virginia public schools, despite evidencing significantly subaverage intellectual functioning and deficits in conceptual, social, and practical adaptive skills. A similar educational history was evidenced in the *Atkins* and *Green v. Virginia* cases.

The fact that none of these individuals had received the label of mental retardation by the public schools in not unusual, particularly for African Americans for whom the issue of overrepresentation in special education programs for the mentally retarded has been an issue since the late 1970s. A study by MacMillan and colleagues showed how this mislabeling" occured in a series of studies conducted in California (Gresham, MacMillan, & Bocian, 1998; MacMillan, Gresham, Bocian, & Siperstein, 1997; MacMillan, Gresham, Siperstein, & Bocian, 1996).

In one study, MacMillan, Gresham, Siperstein, and Bocian (1996) selected a sample of 43 students from grades two, three, and four who had WISC-III IQ scores of 75 and below. The schools that these students attended classified 44% of these students as "learning disabled" (19 students) despite the group having a mean IQ of 68. Only 14% (six students) were classified as mentally retarded with a mean IQ of 63. The remaining 18 students received no formal classification by schools and remained in general education. Similar results were reported by Kanaya, Ceci, and Scullin (2003), who showed that 48.1% of children with IQs below 70 were classified as learning disabled ($M = 66$) and 48.5% were classified as mentally retarded ($M = 64$).

Clearly, relying on a school history of being classified as mentally retarded and receiving special education services under that label is not very reliable in establishing the onset of mental retardation prior to age 18. Courts should be presented with evidence such as that cited by MacMillan, Gresham, Siperstein, and Bocian (1996), MacMillan, Gresham, Bocian, and Siperstein (1997), and Kanaya, Ceci, and Scullin (2003) to demonstrate that the use of the mentally retarded label, especially for individuals with mild mental retardation, is uncommon and is often replaced with a label of learning disabled. Unfortunately, courts often take the failure of schools to diagnose defendants as mentally retarded to be proof that they are not mentally retarded.

CONCLUSIONS

It is clear that experts in *Atkins* cases have provided the court with varying opinions regarding the presence or absence of mental retardation. This was made particularly clear in the *Atkins* trial when one expert diagnosed

Atkins as mentally retarded with an IQ of 59 and the other expert indicated that he had "normal intelligence." An issue that continues to confuse the courts is the nature of *mild mental retardation* (MMR) as distinguished from more severe forms of mental retardation. It is likely that the courts have a preconceived notion of mental retardation that frequently does not include the construct of MMR. Courts are often not convinced that mental retardation, particularly MMR, is a *relative* concept and that an individual's limitations have meaning only in terms of social conditions (Edgerton, 1993). Limitations in intellectual and adaptive behavior functioning must be interpreted within the context of a person's age, culture, and peers and are not absolute concepts. Courts, on the other hand, seek to discover "absolute truths" and are often confounded by arguments that introduce relative concepts into a legal defense.

Differences in expert opinion may stem from a lack of understanding by experts of the concept of mental retardation. Most experts for the prosecution in *Atkins* cases have little or no training or experience in the field of mental retardation (Greenspan, 2006). This was the case in *Walker v. True* in which the expert had a long history of testifying in forensic cases, but no formal training whatsoever in the field of mental retardation.

Another issue that often confuses the courts is the nature of measurement error and how it can affect the interpretation of test scores. This is a nonissue with more severe forms of mental retardation, but a key issue with MMR. If an individual obtains an IQ of 75 and a state uses an IQ below 70 for its intellectual criterion for mental retardation, the prosecution almost always argues that the person cannot be mentally retarded. This argument, however, ignores the fact that there is measurement error in all test scores. For most IQ test scores, the accepted degree of measurement error is five points meaning that an IQ of 75 could be between 70 and 80. More confusing is the fact that measurement error can come from different sources such as internal consistency and stability reliability estimates. In *Walker v. True*, the court considered the concept of measurement error to be "speculative," and defense experts were unsuccessful in arguing against this inaccurate notion.

A controversial issue in *Atkins* cases is the Flynn effect that shows the mean IQ of Americans increases over time by about 0.3 points per year and 3 points per decade (Flynn, 1984). The Flynn Effect can produce a substantial increase in the number of persons diagnosed with MMR, depending on the date a test was normed. For instance, if one used the WISC-III that was normed in 1989 and specified a cutoff of 70 and below, about 2.27% of the population would be identified as mentally retarded. On the other hand, if one used the WISC-IV that was normed in 2001, one would identify approximately 4% of the population as being mentally retarded. The concepts to be understood in interpreting the Flynn Effect are twofold: (1) the *mean* IQ increases over time (the mean shifts upward), and (2) intelligence does not change, only the norms change (i.e., they get "tougher").

Finally, it has been difficult for defendants in *Atkins* cases to meet the developmental criterion in a diagnosis of mental retardation. It must be shown in these cases that an individual's mental retardation had an onset prior to age 18. In many, if not most, *Atkins* cases, this has proven difficult because all defendants have been adults with no prior diagnosis of mental retardation. Consulting defendants' school records frequently show that many of these individuals have a long history of poor academic performance, retention in grade, and a history of special education. In *Atkins, Walker v. True,* and *Green v. Virginia*, all defendants had a history of school difficulties and/or special education, but none were ever diagnosed as being mentally retarded by schools. Instead, Walker was diagnosed as "learning disabled" and "emotionally disturbed" by Richmond, Virginia schools, and Green was diagnosed as "speech and language impaired" and "learning disabled" by the Washington, DC public schools.

It is well-established that schools were and are reluctant to classify children as mentally retarded, particularly African-American students since the 1970s (MacMillan & Siperstein, 2002). Schools frequently assign a more "palatable" label to students who would otherwise be classified as mentally retarded, using labels such as "specific learning disability" or "speech and language impairment." In *Atkins* cases, this frequently works against the defense's efforts because there is no developmental history of an individual ever being diagnosed as mentally retarded, thereby making it difficult to prove the developmental criterion of mental retardation.

Experts testifying for the defense in *Atkins* cases should be well prepared to testify about the nature of mild mental retardation, to be extremely knowledgeable of psychometric theory and measurement error, to understand and be able to articulate the Flynn effect, and to testify about the failure of schools to diagnose mental retardation and their tendency to use "softer" labels for students who may have been mentally retarded. Ultimately, the courts will be the final arbiter of the convincingness of this testimony.

REFERENCES

American Association on Intellectual and Developmental Disabilities. (2002). *Mental retardation: Definition, classification, and systems of supports* (10th ed.). Washington, DC: Author (2002).

American Psychiatric Association. (2000). *Diagnostic and statistical manual for mental disorders* (Text revision). Washington, DC: Author.

Atkins v. Virginia. (2002). 536, U.S. 304, 122, S. CT 2242.

Coombs, C. H. (1950). The concepts of reliability and homogeneity. *Educational and Psychological Measurement*, *10*, 43–56.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Edgerton, R. B. (2003). *The cloak of competence*. Los Angeles: University of California Press.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51.

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*, 170–189.

Green v. Johnson. United States District Court, Eastern District of Virginia, No. 2:05 cv 340. (Testimony in October 2006) Death penalty appeal.

Greenspan, S. (2006). Adaptive behavior in *Atkins* proceedings. Paper presented at a symposium at the Annual Meeting of the American Psychological Association. San Francisco, August 2006.

Gresham, F. M., MacMillan, D., & Bocian, K. (1998). Agreement between school study team decisions and authoritative definitions in classification of students at-risk for mild disabilities. *School Psychology Quarterly*, *13*, 181–191.

Kanaya, T., Ceci, S., & Scullin, M. (2003). The difficulty of basing death penalty eligibility on IQ cutoff scores for mental retardation. *Ethics & Behavior*, *13*, 11–17.

Kaufman, A. (2003). *Practice effects: The clinical café*. Bloomington, MN: Pearson Assessments.

Keith, T. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 581–614). New York: Guilford Press.

MacMillan, D., & Siperstein, G. (2002). Learning disabilities as operationally defined by schools. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Learning disabilities: Research to practice* (pp. 287–333). Mahwah, NJ: Erlbuam.

MacMillan, D., Gresham, F. M., Bocian, K., & Siperstein, G. (1997). The role of assessment in qualifying students as eligible for special education: What is and what's supposed to be. *Focus on Exceptional Children*, *30*, 1–18.

MacMillan, D., Gresham, F. M., Siperstein, G., & Bocian, K. (1996). The labyrinth of IDEA: School decisions on referred students with subaverage general intelligence. *American Journal on Mental Retardation*, *101*, 161–174.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Walker v. True (2006). 399, F. 3d, 327 (4th Cir. 2005).

Wechsler, D. (1972). *Wechsler Memory Scale Manual*. San Antonio: Psychological Corporation.

Standard of Practice and Flynn Effect Testimony in Death Penalty Cases

Frank M. Gresham and Daniel J. Reschly

Abstract

The Flynn Effect is a well-established psychometric fact documenting substantial increases in measured intelligence test performance over time. Flynn's (1984) review of the literature established that Americans gain approximately 0.3 points per year or 3 points per decade in measured intelligence. The accurate assessment and interpretation of intellectual functioning becomes critical in death penalty cases that seek to determine whether an individual meets the criteria for intellectual disability and thereby is ineligible for execution under *Atkins v. Virginia* (2002). We reviewed the literature on the Flynn Effect and demonstrated how failure to adjust intelligence test scores based on this phenomenon invalidates test scores and may be in violation of the *Standards for Educational and Psychological Testing* as well as the "Ethical Principles for Psychologists and Code of Conduct." Application of the Flynn Effect and score adjustments for obsolete norms clearly is supported by science and should be implemented by practicing psychologists.

DOI: 10.1352/1934-9556-49.3.131

The Flynn Effect is a well-established psychometric fact documenting substantial increases in measured intelligence test performance over time. These increases are not generally believed to reflect actual gains in the construct of intelligence but, rather, the creeping obsolesce of test norms (see Flynn, 1984, 1987). Flynn's (1984) seminal review of the literature established that Americans gain an average of approximately 0.3 IQ points per year or 3 points per decade in measured intelligence. His subsequent paper published in 1987 showed a similar increase in measured intelligence worldwide (Flynn, 1987). An intelligence test normed in 1977 and used today has a population mean of approximately 110 (0.3×33 years $= 9.9$). A score of 75 today using the obsolete norms from 1977 is 2.33 *SD* below the population mean and is comparable to a score of 65 if the actual population mean was 100 with an *SD* of 15. The critical issue for psychologists is which score reflects most accurately the individual's current status compared to the overall population.

Our purpose in this article is to provide a discussion of the Flynn Effect and describe how failure to consider it in death penalty cases can have life or death consequences for individuals with intellectual disability. First, we provide an overview of intellectual disability and discuss how so-called *Atkins* cases have exclusively involved individuals having mild intellectual disability rather than more severe forms. We provide a brief overview of relevant aspects of measurement theory and tie this to the legal implications of the Flynn Effect in death penalty cases. We present three actual *Atkins* cases and show how the failure to consider the Flynn Effect, in part, lead to executions in two of the three cases. We conclude the article with a discussion of standards of practice and validity considerations in employing the Flynn Effect in capital cases involving individuals with intellectual disability.

Although widely accepted by scholars, measurement experts, and researchers in the area of intellectual measurement, why, then, is the Flynn Effect important for the everyday practice of clinical assessment? In other words, what practical difference would it make to clinical practitioners

that the population mean changes systematically with the degree of obsolescence of test norms? Moreover, because the scores on tests of intellectual functioning only become meaningful through comparisons to population means, how can clinicians ensure that these comparisons are statistically accurate? Failure to consider changes in measured phenomena or construct over time often can have dire consequences for individuals, and to not account for these changes is to deny this reality.

The accurate assessment of intellectual functioning becomes critical in death penalty cases when determining whether an individual meets the criteria for intellectual disability, in Social Security Administration disability determinations (Reschly, Meyers, & Hartel, 2002), and in eligibility for special education placement and services (MacMillan, Gresham, Siperstein, & Bocian, 1996). In these cases, the use of obsolete norms without appropriate corrections or considerations has enormous consequences for the individual (Flynn, 2010; Flynn & Widaman, 2008). As pointed out by Hagan, Drogin, and Guilmette (2008), psychologists assist in thousands of legal determinations in which the accurate assessment of intellectual functioning is a central issue.

In 2002, the Supreme Court in *Atkins v. Virginia* ruled that it was a violation of the U.S. Constitution Eighth Amendment's prohibition against cruel and unusual punishment to execute individuals with mental retardation. During the Atkins trial, two board certified forensic psychologists came to diametrically opposed opinions concerning whether or not the defendant Daryl Atkins had intellectual disability. One psychologist who evaluated Atkins concluded that he had intellectual disability, with a tested Full-Scale IQ (FSIQ) of 59 on the Wechsler Adult Intelligence Scale-III (WAIS-III). Another forensic psychologist testified that Atkins was functioning in the range of average intelligence. How is it possible that two board certified forensic psychologists can come to vastly different opinions concerning the presence or absence of intellectual disability? As will be illustrated throughout this article, this is neither unexpected nor unusual.

Intellectual Disability

Three prongs have guided the diagnoses of intellectual disability for 70 years (Doll, 1934, 1941): intellectual functioning, adaptive behavior (social competence), and developmental origin. Although classification criteria and terminology differ slightly, *intellectual disability* has been defined by virtually all organizations and states as significantly subaverage intellectual functioning that exists concurrently with deficits in adaptive behavior and which has an onset prior to age 18 years. Most states adopt diagnostic criteria that follow the definition contained in either the *Diagnostic and Statistical Manual (DSM)-TR* (American Psychiatric Association, 2000) or the definition specified by the American Association on Intellectual and Developmental Disabilities—AAIDD (Schalock et al., 2010). Greenspan (2009) has noted that the three criteria specified in the *DSM* and AAIDD manuals have remained conceptually unchanged over nearly 5 decades.

Classification Criteria

What has changed, however, are the operational standards for diagnosing an individual as having intellectual disability based on the criteria of intellectual functioning and adaptive behavior. For example, in the 1961 definition of intellectual disability specified by the American Association on Mental Deficiency—AAMD, Heber (1961) used an intellectual functioning criterion of 85 and below as being indicative of intellectual disability. Twelve years later, the AAMD lowered the intellectual functioning criterion to 70 and below, effectively eliminating 14% of all cases of intellectual disability based on the intellectual functioning criterion (Grossman, 1973).

It is important that both AAIDD and the American Psychiatric Association recognize that measurement error of approximately 5 points is contained in all standardized tests of intelligence and should be taken into account in diagnosing intellectual disability. As such, it is possible to diagnose an individual with intellectual disability who has an IQ up to 75 if they also have significant limitations in adaptive behavior and an onset prior to age 18. One should also realize that there are over twice as many potential cases of intellectual disability with IQs between 70–75 (.0475) than with IQs below 70 (.0222) (Reschly et al., 2002).

The debate in *Atkins* cases has never been about individuals with more severe levels of intellectual disability. It has always been about persons who may be considered to have mild intellectual disability. In the AAIDD *Manual*,

Schalock et al. (2010) defined intellectual disability in much the same way as it was defined in the *DSM-TR* with two exceptions: (a) AAIDD does not specify levels of severity and (b) AAIDD specifies a numerical cutoff score for limitations in adaptive behavior (i.e., greater than 2 *SD*s below the mean) in conceptual, practical, or social adaptive skills.

Types of Intellectual Disability

A crucial issue in *Atkins* cases that is often either misunderstood by the courts or at least is not made clear by defense attorneys is the nature of mild intellectual disability as being distinct from more severe forms. First, mild intellectual disability has no identified or specified biological etiology, whereas more severe forms of intellectual disability often have an identified biological etiology (e.g., Down syndrome, fragile X syndrome, Tay Sachs). Second, mild intellectual disability is most often diagnosed only at school entry or shortly thereafter, whereas severe forms of intellectual disability are often diagnosed at birth or shortly thereafter. Third, some genuine cases of mild intellectual disability are not diagnosed by schools or are misdiagnosed as learning disability (MacMillan et al., 1996). Fourth, adaptive behavior functioning of persons with mild intellectual disability may be adequate in some areas (e.g., practical skills) and severely deficient in others (e.g., conceptual skills). Individuals with severe mental retardation almost always have pervasive deficits in adaptive behavioral functioning. Finally, persons with mild intellectual disability may "blend" into society after school exit (Edgerton, 1993) in that many are not officially diagnosed with intellectual disability in the adult years because they appear to function typically in community settings, whereas persons with severe forms of mental retardation will always "stand out" because of their physical anomalies and severe pervasive intellectual and adaptive behavior deficits. Persons with mild intellectual disability continue, however, to exhibit significant limitations in reasoning and judgment, and the seemingly "normal" performance usually depends on significant assistance from a benefactor (Edgerton, Ballinger, & Herr, 1984).

Many courts may have a preconceived notion of what intellectual disability looks like that is inconsistent with what mild intellectual disability looks like to professionals with training and experience in the field of intellectual disability. Unfortunately, these preconceived notions are often perpetuated by forensic experts who testify for the prosecution and who, more often than not, have little or no training in the field of intellectual disability (Olley, 2009).

Measurement Theory and Intellectual Assessment

A major challenge for any expert witness in *Atkins* cases is to explain to courts the nuances of intellectual assessment and interpretation in understandable terms. Many times, judges, opposing attorneys, and juries have a difficult time understanding how intelligence tests are constructed, what they measure, and how they should be interpreted (Flynn, 2009). For example, in *Atkins* cases, it is important for the court to understand that in a psychometric world, an individual can have more than one true score for his or her level of intellectual functioning. This is particularly true in *Atkins* cases, where defendants often have taken different versions of the same test over time (e.g., the Wechsler scales) and/or different intelligence tests (e.g., Stanford Binet, Woodcock-Johnson, Differential Ability Scales). In many of these cases, an *Atkins* defendant may show higher scores on some intelligence tests and lower scores on others. This is not unusual and can be due to a host of factors, such as different norming periods, different test content, presence or absence of practice effects, and the degree to which the test measures different facets of intelligence (Gresham, 2009).

In classical test theory, an individual's true score on any attribute is entirely dependent on the measurement process that is used (Crocker & Algina, 1986). This is not the case in the biological and physical sciences, in which an individual can have only one true score and that score is independent of the measurement process used. This is known as the *absolute true score*. A relevant example in forensics science is the analysis of a defendant's DNA. Individuals can have only one true score for their DNA, and the courts have come to understand this phenomenon. It is true that different labs may sometimes obtain different results and errors of measurement can occur. This does not alter the fact that only one true score exists for an individual's DNA, and different labs would never average the results of various DNA lab tests to derive a "true DNA score." Yet, this is precisely how we

interpret true scores on psychological measures of intelligence and other attributes.

In classical test theory, an individual can have many true scores for his or her intelligence depending on the number of different intelligence tests administered over his or her lifetime. This logic has been well accepted in the psychometric literature for over 100 years (Spearman, 1904). An *Atkins* case in which we testified brings this interpretative difficulty to light (see *Walker v. True*, 2006). Darick DeMorris Walker was convicted of two capital murders and sentenced to death in Virginia. Walker claimed that the death penalty violated his Eighth Amendment rights to protect him from cruel and unusual punishment because he is mentally retarded. Walker had a history of below-average intellectual functioning and a school history of special education placement. Eventually, Walker dropped out of school in the eighth grade; he had substantial deficits in reading and math skills and a long school history of disruptive and noncompliant behavior.

Seven intelligence tests had been administered to Walker throughout his lifetime, with each test producing somewhat different results. On the various Wechsler tests, Walker's Verbal IQ (VIQ) ranged from 70 to 87, with a median of 78. On the Performance IQ (PIQ) measures, Walker's scores ranged from 61 to 68, with a median of 63. The question before the court in this case was whether or not these scores were indicative of mental retardation. If one takes the VIQ measures at face value, then it is clear that Walker did not meet the Virginia standard for mental retardation. On the other hand, if one takes the various PIQ measures at face value, then it is clear that Walker did meet the Virginia standard for mental retardation. Dilemmas such as these are not uncommon in *Atkins* cases across the country (Greenspan & Switzky, 2006).

In any event, the U.S. District Court (Eastern District of Virginia) ruled against Walker, stating that he failed to show by a preponderance of the evidence that he had intellectual disability. His case was appealed to the U.S. Fourth Circuit Court of Appeals, which vacated and remanded the District Court's judgment and granted Walker an evidentiary hearing to determine whether he had intellectual disability under Virginia law. It further ordered that the District Court should consider all relevant evidence pertaining to Walker's developmental origin, intellectual functioning, and adaptive behavior. The District Court conducted this evidentiary hearing and again reached the conclusion that Walker did not have intellectual disability. Darick Walker was executed by lethal injection at Greensville Correctional Center in Virginia on May 20, 2010.

Legal Implications of the Flynn Effect

There is no doubt that the Flynn Effect can have substantial legal implications in *Atkins* cases in which the presence of intellectual disability for an individual is being contested. As mentioned earlier, in all of these cases, the issue focuses on the category of mild intellectual disability, not more severe cases. Flynn (2006) used the example of a boy who was tested twice during his school years. In 1973, he scored 75 on the WISC that was normed in 1947–1948; thus, the norms were 25.5 years out of date. In 1975, the boy was tested at age 8 with the WISC-R, which was normed in 1972, and, therefore, with norms only 3 years out of date. He obtained an IQ of 68. The score at age 6 of 75 and at age 8 of 68 are, in fact, statistically the same score based on the Flynn Effect because the 1973 score was inflated by 7 points and the 1975 score was not influenced by the Flynn Effect because of the recency of the WISC-R norms.

How is this example relevant to present day *Atkins* cases? Suppose two defendants were tested in 2004 to provide evidence that would be presented in *Atkins* cases. The first defendant was tested with the WAIS-III that was normed in 1989 and obtained an IQ of 73. The second defendant was tested with the WAIS-IV that was normed in 2002 and obtained a score of 69. The first defendant was convicted and sentenced to death because his score did not meet the "bright line" of IQ 70 or below, whereas the second defendant was not sentenced to death because his IQ of 69 met the state's bright line of IQ less than 70. The fact is that both of these scores for the two defendants are statistically identical when viewed in light of the Flynn Effect.

This is precisely what happened in a recent Florida *Atkins* case (Cherry v. State, 2007). Roger Cherry was convicted of capital murder and sentenced to death. On a postconviction appeal, Cherry claimed he had intellectual disability and, therefore, was ineligible for the death penalty. His tested WAIS-III score of 72 did not meet the Florida bright line criterion of IQ 70 and below, and the court denied Cherry's appeal. In fact, when

Cherry took the WAIS-III, the norms were 13 years out of date, thereby producing a Flynn Effect of approximately 4 points. Based on the Flynn Effect, Cherry's IQ of 72 is actually 68, thereby meeting the Florida bright line standard. As Flynn (2006) indicated: "Failure to adjust IQ scores in light of IQ gains over time turns eligibility for execution into a lottery" (pp. 174–175).

Some of the illustrations above might be criticized because they are hypothetical; however, we next present three actual *Atkins* cases that show the real legal ramifications of the Flynn Effect in death penalty cases. The first case presented in Table 1 is Darick Walker (previously mentioned), who was convicted of two capital murders (*Walker v. True, 2006*) and executed on May 20, 2010. Recall that the U.S. District Court ruled twice that Walker did not have intellectual disability and upheld his death penalty sentence. Table 1 shows that Walker's Wechsler IQs for VIQ, PIQ, and FSIQ were 70, 85, and 76, respectively. When Flynn corrections were applied, these scores more accurately were 66, 81, and 72, respectively, and clearly placed Walker in the range of mild intellectual disability based on *DSM-TR* and AAIDD intellectual criteria.

The second case presented in Table 1 is Kevin Green, who was convicted of capital murder, denied a status of mental retardation in an appeal of the death penalty (*Green v. Johnson, 2006*), sentenced to death, and executed on May 27, 2008. Green's IQs were 67, 80, and 71 for VIQ, PIQ, and FSIQ, respectively. In 1991, while a 14-year-old student in fourth grade (having failed three school grades previously and described by his teacher as fitting in well socially with children 4 to 5 years younger), Green was referred for a psychological evaluation as part of the consideration of special education eligibility. The 1974 version of the Wechsler Scale (WISC-R) was used, despite the publication of the updated WISC-III in 1991. The FSIQ of 71 was derived from a test with norms that were 19 years obsolete. The WISC-R population mean in 1991 was approximately 106. The score of 71 on the WISC-R in 1991 was 2.33 *SD*s below the population mean, clearly exceeding the traditional standard of intellectual functioning approximately 2 *SD* below the population mean. However, the Flynn corrections show that Green's scores in comparison to the existing population mean were 61, 74, and 65, respectively, clearly placing him in the range of mild intellectual disability based on the intellectual criterion. Nevertheless, a board certified forensic psychologist urged the court to ignore the Flynn Effect because it did not represent the current standard of practice in psychology (see later discussion).

Table 1 Uncorrected and Flynn Corrected Wechsler Scores for Three *Atkins* Cases

Score[a]	Walker[b]	Green[c]	Johnston[d]
VIQ	70	67	69
FVIQ	66	61	63
PIQ	85	80	89
FPIQ	81	74	83
FSIQ	76	71	76
FFSIQ	72	65	71

[a]VIQ = Verbal IQ, FVIQ = Flynn Corrected VIQ, PIQ = Performance IQ, FPIQ–Flynn Corrected PIQ, FSIQ=Full Scale IQ, FFSIQ–Flynn Corrected FSIQ. [b]Based on WAIS-III normed in 1989 and administered in 2004. [c]Based on WISC-R normed in 1972 and administered in 1991. [d]Based on WAIS-III normed in 1989 and administered in 2005.

Finally, Table 1 shows the Wechsler IQs for David Johnston, who was convicted of capital murder in Florida (see *Johnston v. State*, 1986) and sentenced to death. Table 1 shows that Johnston's IQs were 69, 89, and 76 for VIQ, PIQ, and FSIQ, respectively. Flynn corrections lower these scores to 63, 83, and 70, respectively, again placing Johnston in the range of mild intellectual disability based on the intellectual criterion.

All three of the above cases consistently show how failure to account for the Flynn Effect can produce IQs that move defendants out of the range of intellectual disability on the Wechsler scales. In 2 of the 3 cases (Walker and Green), this failure contributed to their execution in the state of Virginia. The third case (Johnston) was before the Florida Supreme Court; however, Johnston died of natural causes on Death Row before the Supreme Court could rule on his case.

Some have questioned whether or not the Flynn Effect applies reliably to specific individuals, particularly those who find themselves in *Atkins* cases and death penalty appeals (Hagan et al., 2008). This is, frankly, a specious argument simply because any individual's IQ is entirely dependent upon group mean scores of the standardization sample. If the group mean has shifted upward, then the score that meets the intellectual disability

standard has likewise increased by the same amount (Flynn, 1985). If this standardization sample is obsolete, then any individual score calculated in reference to the obsolete norms will be inflated by a factor of 0.3 points per year, or 3 points per decade from when the test was standardized.

The Flynn Effect has a substantial influence on the number of persons who might be classified as having intellectual disability using a specified cutoff score based on a large scale of the proportions of persons identified as having intellectual disability and placed in special education programs. For example, Kanaya, Ceci, and Scullin (2003) found that the number of children who were diagnosed with intellectual disability nearly tripled with the introduction of the WISC-III (from the WISC-R) because more and more children obtained an IQ of 70 and below with the comparison to the more difficult norm. The Flynn Effect produces situations in which a given individual's IQ can fluctuate above and below a specified IQ cutoff that most states use to determine eligibility for the death penalty (Flynn, 2009; Kanaya et al., 2003). In effect, this is like playing dice with IQ scores, except the stakes in *Atkins* cases are most certainly higher.

Two recent court cases in capital trials applied the Flynn Effect as well as acknowledging the standard error of measurement and an intellectual disability cutoff score at 75 to evidence similar to that in the *Walker* and *Green* cases, leading to decisions forbidding the death penalty (*U.S. v. Hardy*, 2010; *U.S. v. Lewis*, 2010). It is significant that these cases were trials in federal district courts, where the judges are appointed for life, rather than in state courts, where judges often are elected and more responsive to public opinion, which frequently favors strong retribution against capital defendants. In both of the recent cases, the Flynn Effect was accepted as a scientific fact, and testimony that the Flynn Effect is not currently taught in graduate programs preparing psychologists was essentially discounted. We can only speculate on whether state courts will increasingly adopt what we see as clear scientific evidence cases confirming the Flynn Effect.

We acknowledge that acceptance of the Flynn Effect will not always yield decisions forbidding the death penalty. In fact, in both *Green* and *Walker*, the appellants were also found ineligible for the intellectual disability classification on the adaptive behavior criterion. It is our impression, however, that courts, much like practitioners making diagnoses of intellectual disability in school settings, are strongly influenced by the individual's status on the general intellectual functioning prong, with decisions about adaptive behavior following rather than being equally weighted with intelligence in intellectual disability decisions (Reschly & Ward, 1991). Greater weighting of the intellectual prong also occurs because of less well-developed measures of adaptive behavior and difficulties with gathering adaptive behavior information for adults prior to age 18 (Reschly, 2009).

Standard of Practice and the Flynn Effect

What, then, are practicing psychologists to do when presented with an *Atkins* case, and they find themselves as expert witnesses in courts or in SSI disability evaluations involving intellectual disability? In other words, what is the appropriate standard of practice for interpreting IQs in light of the Flynn Effect? Opinions regarding this issue understandably vary depending on who is asked that question. Greenspan (2006) suggested that adjusting an individual's IQ in light of the Flynn Effect is essential. Others have made similar suggestions based on their analysis of the Flynn Effect in various reviews of the literature (Ceci & Kanaya, 2010; Fletcher, Stuebing, & Hughes, 2010; Kanaya et al., 2003; McGrew, 2010).

Hagan et al. (2008) addressed this issue by conducting a survey of 358 APA-approved clinical, counseling, and school psychology program directors. One surprising result was the fact that over one third (36%) of program directors had either not heard of the Flynn Effect or were slightly familiar with the concept. Of the remaining 64% of the respondents, almost 92% of them indicated they would never teach students to recalculate IQs based on the Flynn Effect. Similarly, a survey of 28 Diplomates in School Psychology revealed that 94% of them had never adjusted IQs based on the Flynn Effect.

Survey results depend heavily on how questions are worded and the use of context descriptions. Apparently, Hagan et al. (2008) simply inquired about subtracting points based on the Flynn Effect without any description of context or implications. Under these circumstances the clear majority of the small proportions of each sample who responded rejected score adjustments. These results likely would have been different if the respondents were given SSI or death penalty contexts, such as those described above in the Walker, Green, and Johnston cases.

Hagan et al. (2008) also reported that primary source assessment texts and test manuals did not recommend changing scores. Again, however, context and vested interests likely make a difference. Moreover, test publishers have a vested interest in ignoring the Flynn Effect in test manuals because of the tacit admission attendant to discussing this phenomenon that tests have a limited shelf life and need to be updated frequently (Kaufman, 2010; Weiss, 2007, 2010). One exception is the following content from the WAIS-III *Manual* (Wechsler, 1997).

> **Updating of Norms**. Because there is a real phenomenon of IQ-score inflation over time, norms for a test of intellectual functioning should be updated regularly (Flynn 1984, 1987; Matarazzo, 1972). Data suggest that an examinee's IQ score will generally be higher when outdated rather than current norms are used. The inflation rate of IQ scores is about 0.3 points each year. Therefore, if the mean IQ of the U.S. population on the WAIS-R was 100 in 1981, the inflation might cause it to be about 105 in 1997. (pp. 8–9)

Not surprisingly, the most recent WAIS version does not discuss the Flynn Effect (Wechsler, 2008), perhaps reflecting the rather defensive denial of Flynn's criticism of the WAIS-III standardization sample by a test company official involved with the development of the Wechsler scales (Weiss, 2007). To set the record straight, the Flynn Effect continues to be prominent and well supported statistically through the most recent revisions of the Wechsler scales (Flynn, 2009).

Hagan et al. (2008) concluded that adjusting IQ scores and recalculating scores based on the Flynn Effect do not represent custom or standard of practice in professional psychology based on a survey with a participation rate among those surveyed. This so-called standard of practice, however, was based on a survey in which over one third of the sample responding was fundamentally unfamiliar with the concept at issue—namely, the Flynn Effect. The majority of the remaining respondents said they would never teach students to adjust scores based on the Flynn Effect. This finding is not scientifically convincing and should not be taken at face value. The Flynn Effect is a well-established measurement phenomenon based on years of replicated research findings across the world. The fact that most program directors would never teach students to interpret scores in light of the Flynn Effect is to ignore scientific reality and potentially could be in violation of the *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999).

Perhaps the most well-known and qualified group of professionals who deal with the diagnosis and treatment of persons with intellectual disability are members of the AAIDD. Founded in 1876, this organization has, through 11 editions of its diagnostic manual, provided guidance for professionals working in the field of intellectual disability. Reschly (1992) established that the AAIDD leads the world, including the *DSM*, in the development and refinement of the intellectual disability diagnostic construct. In the User's Guide of the 10th edition of the AAIDD Manual, Schalock et al. (2006) stated that best practices require recognition of the Flynn Effect when older editions of an intelligence test are used in assessment or interpretation of an IQ score. The authors go further:

> The main recommendation resulting from this work [regarding the Flynn Effect] is that all intellectual assessment must use a reliable and appropriate individually administered intelligence test. In cases with multiple versions, the most recent version with the most current norms should be used at all times. *In cases where a test with aging norms is used, a correction for the age of the norms is warranted* [italics added]. (pp. 20, 21)

Validity Considerations

Validity is the centerpiece concept in every aspect of psychological assessment. Validity is an evaluative judgment of the extent to which empirical evidence and theoretical explanations support the adequacy and appropriateness of test score interpretations and actions (Messick, 1995). We emphasize that validity is not a characteristic of a given test, but rather is a property of the meaning of test scores. Cronbach (1971) argued that what is validated in psychological testing is the meaning and interpretation of the test score and the implications for actions that the meaning entails.

Based on this conceptualization of validity, what impact does the Flynn Effect have on the meaning and interpretation of intelligence test scores? The most obvious implication is that failure to account for the Flynn Effect in the interpretation of such scores renders that interpretation inaccurate. For example, interpretation of a WAIS-III score of 72 administered in 2006 and deciding that the individual does not meet the criterion of IQ 70 or less would be erroneous. A Flynn correction of this score, in fact, would yield a more accurate score of 69, thereby meeting the IQ criterion. It is unknown how prevalent these validity violations are in *Atkins* cases, but we believe this to be a

common phenomenon, particularly based on the Hagan et al. (2008) survey of clinical, counseling, and school psychology program directors.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) indicate that proper interpretations of test scores may be compromised by *construct-irrelevant variance*, which is defined as the degree to which test scores are affected by processes that are extraneous to the construct being measured. We argue that the failure to adjust IQ scores based on the Flynn Effect introduces construct-irrelevant variance into the proper interpretation of intelligence test scores. Failure to make this adjustment diminishes the quality and accuracy of test score interpretation and invalidates the inferences that can be made from those test scores.

Messick (1995) discussed the issue of consequential validity in his seminal paper on validity of psychological assessment. Using the language of Cronbach and Meehl (1955), Messick suggested that unintended consequences occurring in psychological testing are strands in the nomological network that should be taken into account in test score interpretation and use. We maintain that failure to account for the Flynn Effect in death penalty cases can produce adverse social consequences for individuals and, thus, invalidate their test scores. Messick (1995) suggested that:

> The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence. (p. 746)

We argue that this same logic also works in the opposite direction. That is, higher scores should not occur because the measurement contains something irrelevant that interferes with an affected person's demonstration of lowered intellectual functioning. The Flynn Effect injects such construct irrelevant variance into the interpretation of test scores when professional psychologists do not account for it.

The Flynn Effect and its proper use in professional psychological practice might be cast in terms of the value implications to proper test score interpretation. Value implications are an integral aspect of proper test score interpretation and often link the construct being assessed to questions of applied practice and social policy (Messick, 1995). The proper use of the Flynn Effect in *Atkins* cases, we think, captures the essence of what Messick meant by value implications and proper test score interpretation. To this we would add that Principle 9.08 (Obsolete Tests and Outdated Test Results) of the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002) states in part: "*(B) Psychologists do not base such decisions or recommendations on tests and measures that are obsolete and not useful for the current purpose* [italics added]." Failure to account for the Flynn Effect in test score interpretation in *Atkins* or any other cases is a violation of this ethical principle. In addition, failure to ensure the accurate interpretation of test scores in *Atkins* cases may possibly be a violation of the ethical Principle A: Beneficence and Nonmaleficence of the APA Code of Ethics. The principle states, in part, "Psychologists strive to benefit those with whom they work and take care to *do no harm* [italics added]." In their professional actions, psychologists seek to safeguard the welfare and rights of those with whom they interact professionally and other affected persons.

Given that *Atkins* held that it is a violation of the Eighth Amendment to the Constitution to execute persons who suffer from intellectual disability, it would seem that concluding individuals do not have intellectual disability without considering the Flynn Effect most certainly would cause undue harm and would violate the Constitutional rights of these individuals.

Conclusion

Standard of practice in the use of the Flynn Effect in the context of high stakes decisions must be guided by scientific evidence, not by opinion of psychologists. As Hagen et al. (2008) found in their survey, many psychologists are not aware of the underlying science and likely not cognizant of the high stakes contexts. Practicing psychologists claim to use an underlying psychological science as the foundation for clinical work. Application of the Flynn Effect and score adjustments for obsolete norms clearly is supported by science and should be implemented by professional psychologists.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text. rev.). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073.

Atkins v. Virginia. 536, U.S. 304, 122, S. CT 2242. (2002).

Ceci, S. J., & Kanaya, T. (2010). "Apples and oranges are both round": Furthering the discussion of the Flynn Effect. *Journal of Psychological Assessment, 28*, 441–447.

Cherry v. State, 959 So. 2d 702, 712-13 (Fla. 2007)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Doll, E. A. (1934). Social adjustment of the mental subnormal. *Journal of Educational Research*, *28*, 36–43.

Doll, E. A. (1941). The essentials of an inclusive concept of mental deficiency. *American Journal of Mental Deficiency*, *46*, 214–219.

Edgerton, R. B. (1993). *The clock of competence: Revised and updated*. Berkeley: University of California Press.

Edgerton, R. B., Ballinger, M., & Herr, B. (1984). The cloak of competence: After two decades. *American Journal of Mental Deficiency*, *88*, 345–351.

Fletcher, J. M., Stuebing, K. K., & Hughes, L. C. (2010). IQ scores should be corrected for the Flynn Effect in high-stakes decisions. *Journal of Psychoeducational Assessment, 28*, 469–473.

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal on Mental Deficiency, 90*, 236–244.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Flynn, J. R. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn Effect. *Psychology, Public Policy, and Law*, *12*, 170–189.

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: *Daubert* motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*, 98–104.

Flynn, J. R. (2010). Problems with IQ gains: The huge vocabulary gap. *Journal of Psychoeducational Assessment, 28*, 412–433.

Flynn, J. R., & Widaman, K. F. (2008). The Flynn Effect and the shadow of the past: Mental retardation and the indefensible and indispensable role of IQ. *International Review of Research in Mental Retardation, 35*, 121–149.

Green v. Johnson, 2006 U.S. Dist. LEXIS 90644 (E.D. Va.), adopted by, 2007 U.S. Dist. LEXIS 21711 (E.D. Va.), aff'd., 2008 U.S. App. LEXIS 2967 (4th Cir.), cert. denied, 128 S. Ct. 2527 (2008).

Greenspan, S. (2006, spring). Issues in the use of the "Flynn Effect" to adjust IQ scores when diagnosing MR. *Psychology in Mental Retardation and Developmental Disabilities, 31*, 3–7.

Greenspan, S. (2009). Assessment and diagnosis of mental retardation in death penalty cases: Introduction and overview of the special "Atkins" issue. *Journal of Psychoeducational Assessment, 16*, 89–90.

Greenspan, S., & Switzky, H. (2006). Lessons from the *Atkins* decision for the next AAMR manual. In H. Switzky & S. Greenspan (Eds.), *What is mental retardation? Ideas for an evolving disability in the 21st century* (pp. 281–300). Washington, DC: American Association on Mental Retardation.

Gresham, F. M. (2009). Interpretation of intelligence test scores in *Atkins* cases: Conceptual and psychometric issues. *Applied Neuropsychology, 16*, 91–97.

Grossman, H. J. (Ed.). (1973). *Manual on terminology and classification in mental retardation*. Washington, DC: American Association on Mental Deficiency.

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2008). Adjusting IQ scores for the Flynn Effect: Consistent with standard of practice? *Professional Psychology: Research and Practice, 39*, 619–625.

Heber, R. (1961). *A manual on terminology and classification in mental retardation* (Rev. ed.). Washington, DC: American Association on Mental Deficiency.

Johnston v. State. 497 So. 2d 863 (Fla.1986).

Kanaya, T., Ceci, S., & Scullin, M. (2003). The Flynn Effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist, 58*, 778–790.

Kaufman, A. S. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn Effect. *Journal of Psychoeducational Assessment, 28*, 382–398.

MacMillan, D., Gresham, F. M., Siperstein, G., & Bocian, K. (1996). The labyrinth of IDEA: School decisions on referred students with subaverage general intelligence. *American Journal on Mental Retardation, 101*, 161–174.

Matarazzo, D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th enlarged ed.). Baltimore: Williams & Wilkins.

McGrew, K. S. (2010). The Flynn Effect and its critics: Rusty linchpins and "lookin' for g and Gf in some of the wrong places." *Journal of Psychoeducational Assessment, 28*, 448–468.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Olley, J. G. (2009). Knowledge and experience required for experts in *Atkins* cases. *Applied Neuropsychology, 16*, 135–140.

Reschly, D. J. (1992). Mental retardation: Conceptual foundations, definitional criteria, and diagnostic operations. In S. R. Hooper, G. W. Hynd, & R. E. Mattison (Eds.), *Developmental disorders: Diagnostic criteria and clinical assessment* (pp. 23–67). Hillsdale, NJ: Erlbaum.

Reschly, D. J. (2009). Documenting the developmental origins of mild mental retardation. *Applied Neuropsychology, 16*, 124–134.

Reschly, D. J., Myers, T. G., & Hartel, C. R. (Eds.). (2002). *Mental retardation: Determining eligibility for Social Security benefits*. Washington, DC: National Academy Press.

Reschly, D. J., & Ward, S. M. (1991). Use of adaptive measures and overrepresentation of black students in programs for students with mild mental retardation. *American Journal of Mental Retardation*, 96, 257–268.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M., et al. (2010). *Intellectual disability: Definition, classification, and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Schalock, R., Buntinx, W., Borthwick-Duffy, Luckasson, R., Snell, M., Tassé, M., & Wehmeyer, M. (2006). *User's guide: Mental retardation, classification, and systems of supports,10th Edition: Applications for clinicians, educators, disability program managers, and policy makers*. Washington, DC: American Association on Intellectual and Developmental Disabilities.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

United States v. Hardy (2010, November 24). U.S. District Court Eastern District of Louisiana, CA No. 94–381.

United States v. Lewis (2010, December 23). Case No.: 1:08 CR 404. U.S. District Court Northern District of Ohio Eastern Division. (Trial Opinion)

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corp.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio, TX: Psychological Corp.

Weiss, L. G. (2007). *WAIS-III: Technical report, Response to Flynn*. San Antonio, TX: Psychological Corp.

Weiss, L. G. (2010). Considerations on the Flynn Effect. *Journal of Psychoeducational Assessment, 28*, 482–493.

Walker v. True (2006, August 30). U.S. District Court for the Eastern District of Virginia, Alexandria Division, Case No. 1:03–cv–00764.

Walker v. True (2006). 399 F. 3d, 327 (4th cir, 2005).

Received 10/26/10, first decision 1/19/11, accepted 1/31/11.

Editor-in-Charge: Steven J. Taylor

Authors:
Frank M. Gresham, PhD (e-mail: frankgresham@yahoo.com), Professor, Department of Psychology, Louisiana State University, Baton Rouge, LA 70803.
Daniel J. Reschly, PhD, Professor, Department of Special Education, Vanderbilt University, Nashville, TN 37203.