Transparency and Fairness in Algorithms for Criminal Justice

Cristopher Moore, Santa Fe Institute
Kathy Powers, UNM Political Science
Interdisciplinary Working Group
on Algorithmic Justice

Interdisciplinary Working Group on Algorithmic Justice

Cris Moore
Santa Fe Institute

Kathy Powers
Political Science

Melanie Moses
Computer Science

Alfred Mathewson
Law

Sonia Rankin
Law

Mirta Galesic
Santa Fe Institute

Josh Garland
Santa Fe Institute

Matthew Fricke
Computer Science

Gabe Sanchez
Political Science

Interdisciplinary Working Group on Algorithmic Justice

Who are we?

Independent scientists and legal scholars

University of New Mexico: Computer Science, Political Science, Law

Santa Fe Institute: Computer Science, Applied Mathematics, Statistics, Social Psychology

What are our goals?

To act as a resource to policymakers and stakeholders

To *demystify* algorithms, and explain their strengths and weaknesses

To offer policy advice about if, when, and how algorithms should be deployed in the public sector

Algorithms and Justice

Used increasingly for high-stakes decisions affecting lives and liberties:

- Housing and lending: mortgages, loans, rentals
- Policing: predicting crime, identifying subjects
- Social services, child protective services
- Criminal justice
  - Pretrial supervision and detention
  - Sentencing
  - Housing classification in prison
  - Parole

Algorithms and Justice

What is an algorithm? (a.k.a. risk assessment instruments, actuarial tools)

- It takes input about a defendant (e.g. their criminal record)
- Based on statistical patterns★ in a database of past cases (the “training data”)
- …and the assumption that this defendant will have similar outcomes to defendants in the training data with similar records…
- …the algorithm estimates the risk (probability) that this defendant will have outcomes such as:
    - Failure To Appear: missing one or more court hearing
    - New Criminal Activity: arrested for new offense while awaiting trial
    - Recidivism (for parole), infractions (for prisoners), etc.

★ human choices: what data to collect, what kind of patterns to look for

Algorithms and Justice

Claim by the proponents: algorithms are more accurate, less biased, more objective than humans. This may or may not be true!

But what kind of **transparency** do we need to ensure that these algorithms are accurate and fair? Some good questions:

1. How does the algorithm work? Can everyone (defendants, prosecutors, judges) understand how a score was obtained?

2. Can we validate its performance independently? How well does it work on our local population in New Mexico?

3. When should a human be in the loop? Should an algorithm ever be used for detention before trial?

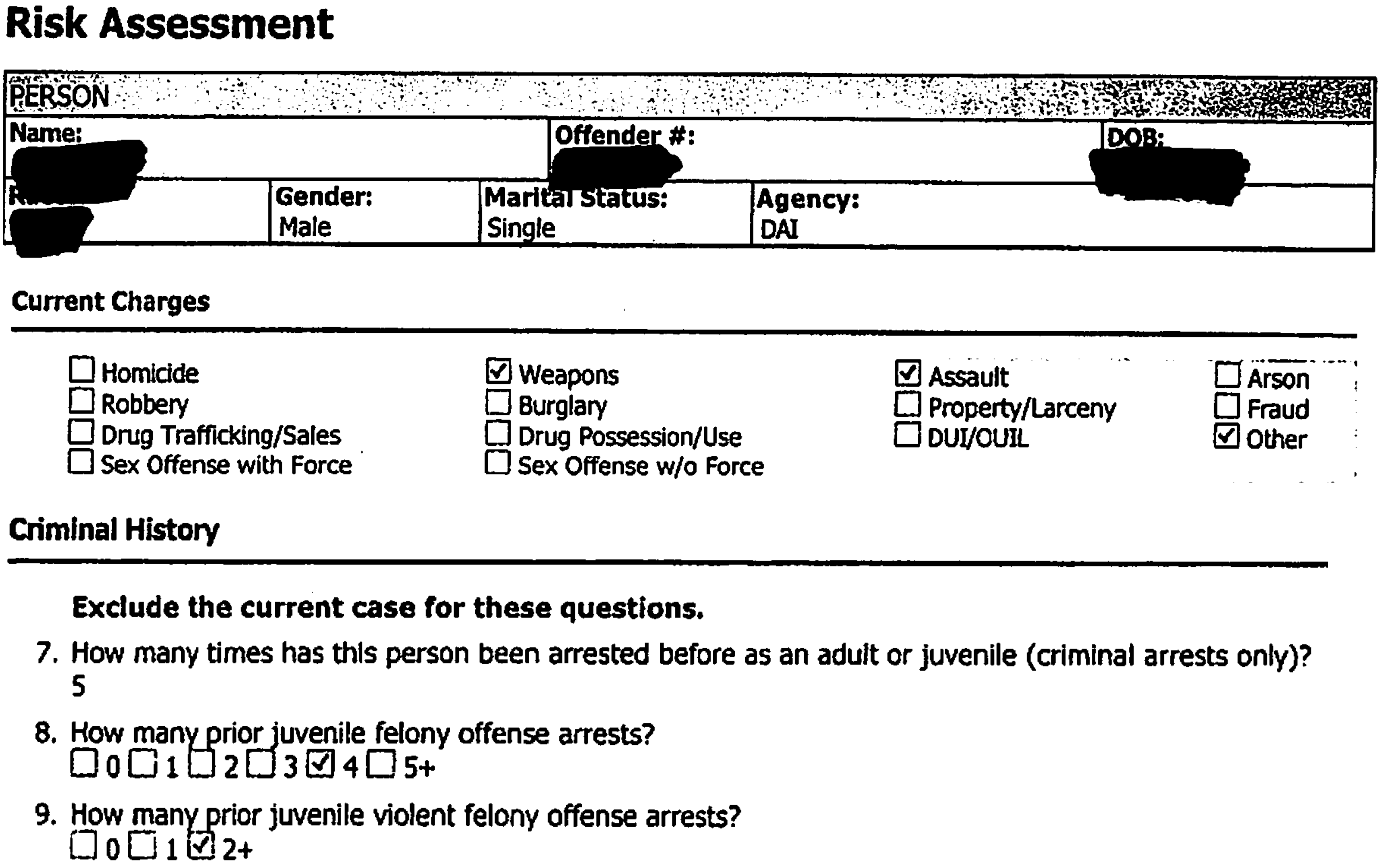
4. What does the data really mean? Does a single zero or one capture the full story behind a failure to appear or rearrest?

Transparency #1: How Does the Algorithm Work?

Two popular algorithms at opposite ends of the transparency spectrum

- COMPAS
  - Northpointe / equivant
  - 137-item questionnaire and interview
  - Proprietary (secret) formula
- Arnold Public Safety Assessment (PSA)
  - Rapidly growing, four states and 40 jurisdictions
  - 9 factors from criminal record
  - Simple, transparent formula

What data goes into COMPAS?



Risk Assessment

PERSON

Name: [redacted]

Offender #: [redacted]

DOB: [redacted]

[redacted]

Gender: Male

Marital Status: Single

Agency: DAI

Current Charges

□ Homicide
□ Robbery
□ Drug Trafficking/Sales
□ Sex Offense with Force

☑ Weapons
□ Burglary
□ Drug Possession/Use
□ Sex Offense w/o Force

☑ Assault
□ Property/Larceny
□ DUI/OUIL

□ Arson
□ Fraud
☑ Other

Criminal History

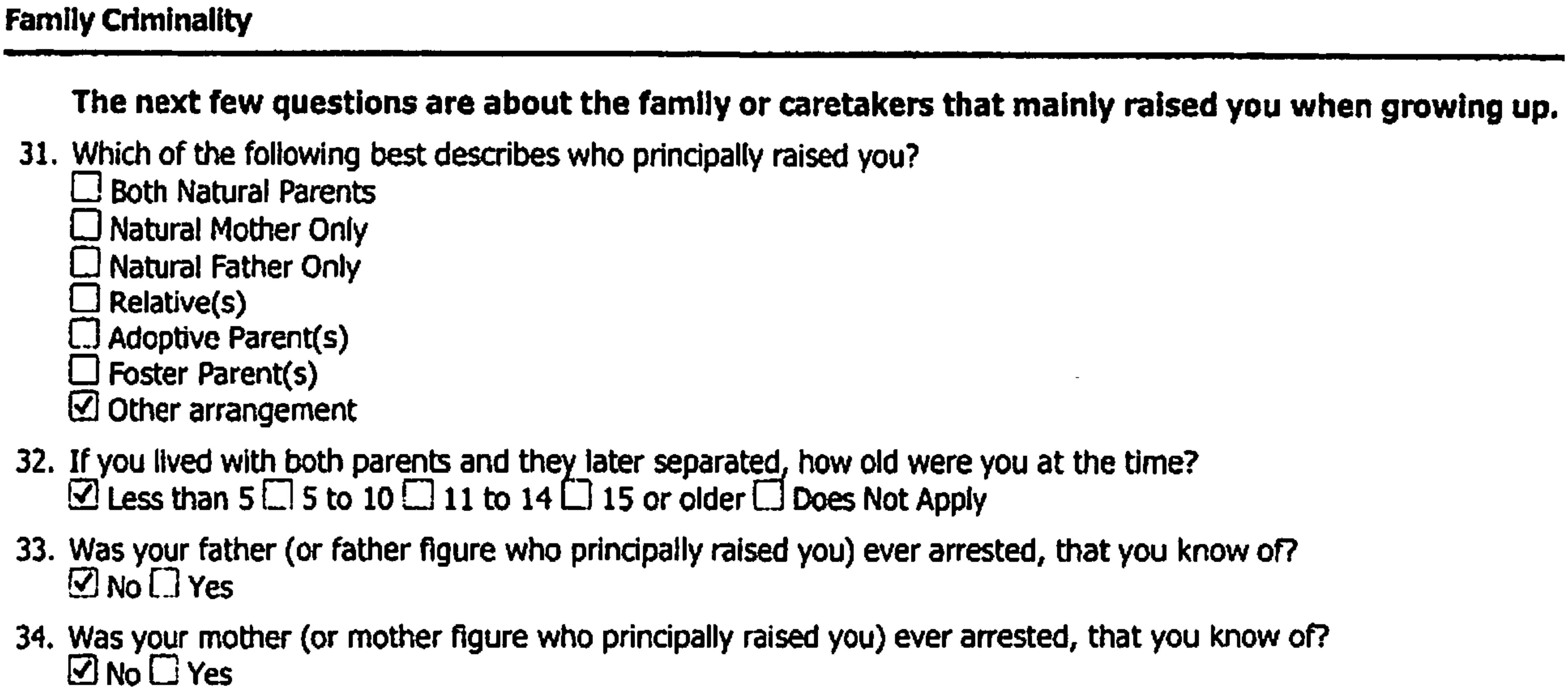
Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
5

8. How many prior juvenile felony offense arrests?
□ 0 □ 1 □ 2 □ 3 ☑ 4 □ 5+

9. How many prior juvenile violent felony offense arrests?
□ 0 □ 1 ☑ 2+

What data goes into COMPAS?



Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

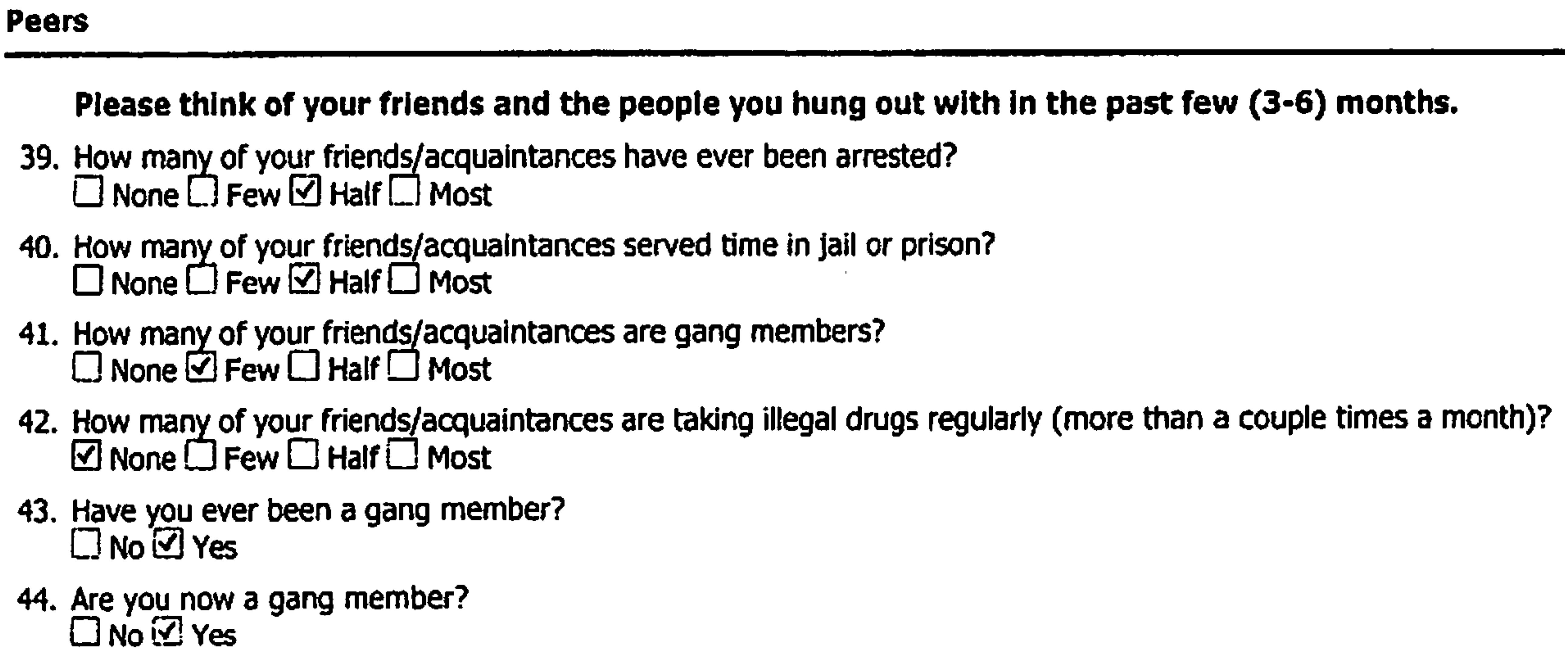
31. Which of the following best describes who principally raised you?
    - □ Both Natural Parents
    - □ Natural Mother Only
    - □ Natural Father Only
    - □ Relative(s)
    - □ Adoptive Parent(s)
    - □ Foster Parent(s)
    - ☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
    ☑ Less than 5 □ 5 to 10 □ 11 to 14 □ 15 or older □ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
    ☑ No □ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
    ☑ No □ Yes

What data goes into COMPAS?



Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
☐ None ☐ Few ☑ Half ☐ Most

40. How many of your friends/acquaintances served time in jail or prison?
☐ None ☐ Few ☑ Half ☐ Most

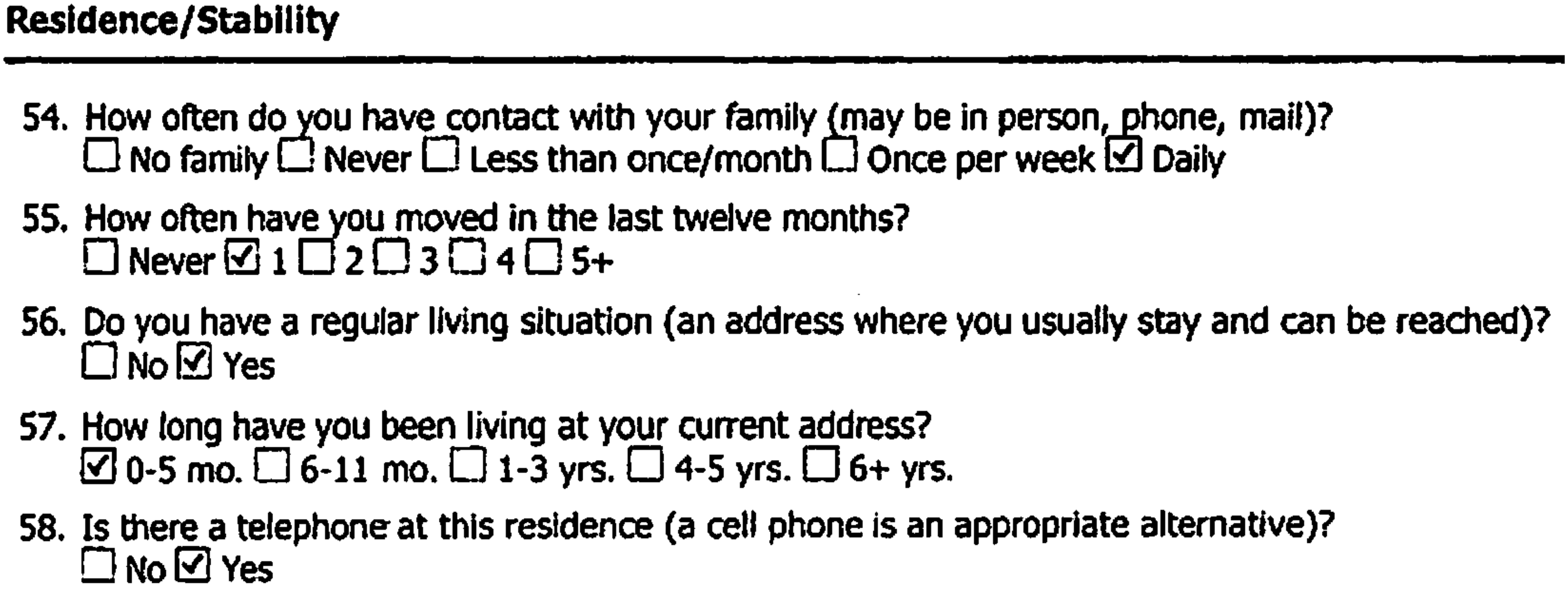
41. How many of your friends/acquaintances are gang members?
☐ None ☑ Few ☐ Half ☐ Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?
☑ None ☐ Few ☐ Half ☐ Most

43. Have you ever been a gang member?
☐ No ☑ Yes

44. Are you now a gang member?
☐ No ☑ Yes

What data goes into COMPAS?



Residence/Stability

54. How often do you have contact with your family (may be in person, phone, mail)?
□ No family □ Never □ Less than once/month □ Once per week ☑ Daily

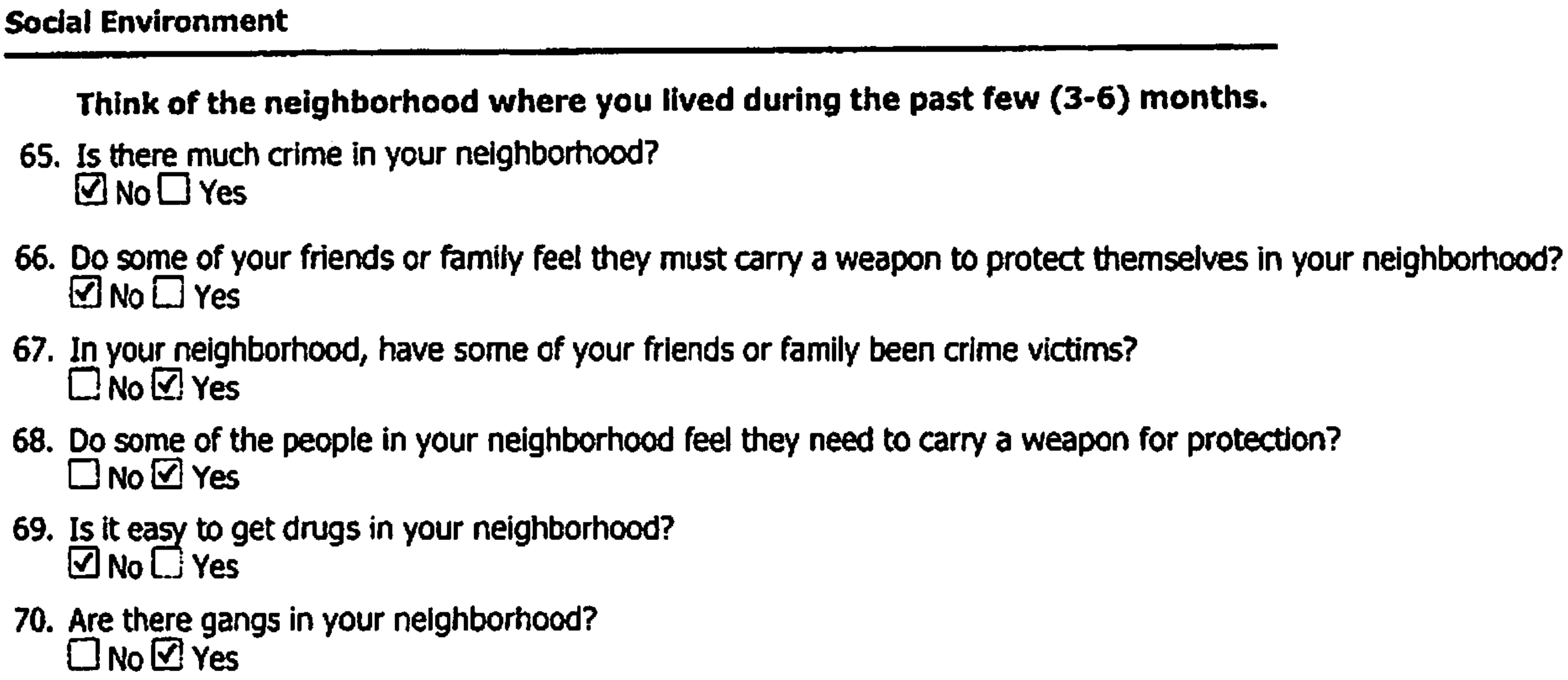
55. How often have you moved in the last twelve months?
□ Never ☑ 1 □ 2 □ 3 □ 4 □ 5+

56. Do you have a regular living situation (an address where you usually stay and can be reached)?
□ No ☑ Yes

57. How long have you been living at your current address?
☑ 0-5 mo. □ 6-11 mo. □ 1-3 yrs. □ 4-5 yrs. □ 6+ yrs.

58. Is there a telephone at this residence (a cell phone is an appropriate alternative)?
□ No ☑ Yes

What data goes into COMPAS?



Social Environment

Think of the neighborhood where you lived during the past few (3-6) months.

65. Is there much crime in your neighborhood?
☑ No □ Yes

66. Do some of your friends or family feel they must carry a weapon to protect themselves in your neighborhood?
☑ No □ Yes

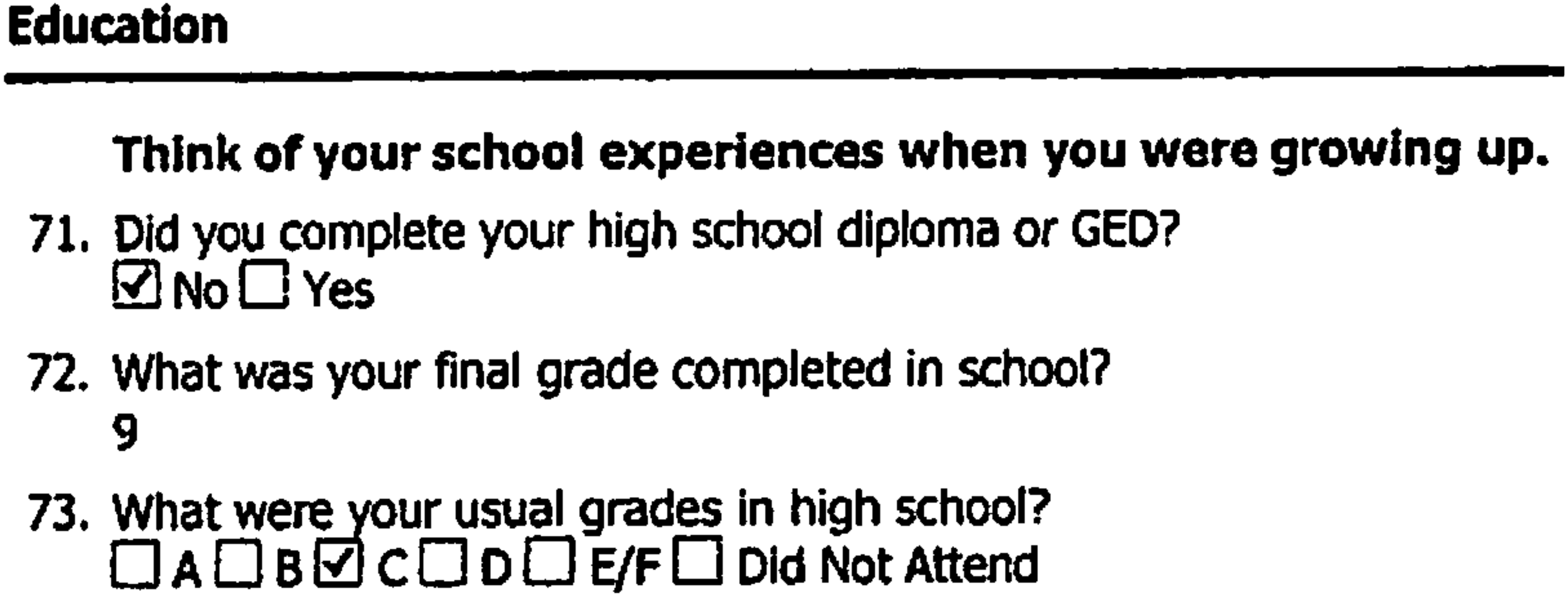
67. In your neighborhood, have some of your friends or family been crime victims?
□ No ☑ Yes

68. Do some of the people in your neighborhood feel they need to carry a weapon for protection?
□ No ☑ Yes

69. Is it easy to get drugs in your neighborhood?
☑ No □ Yes

70. Are there gangs in your neighborhood?
□ No ☑ Yes

What data goes into COMPAS?



Education

Think of your school experiences when you were growing up.

71. Did you complete your high school diploma or GED?
☑ No □ Yes

72. What was your final grade completed in school?
9

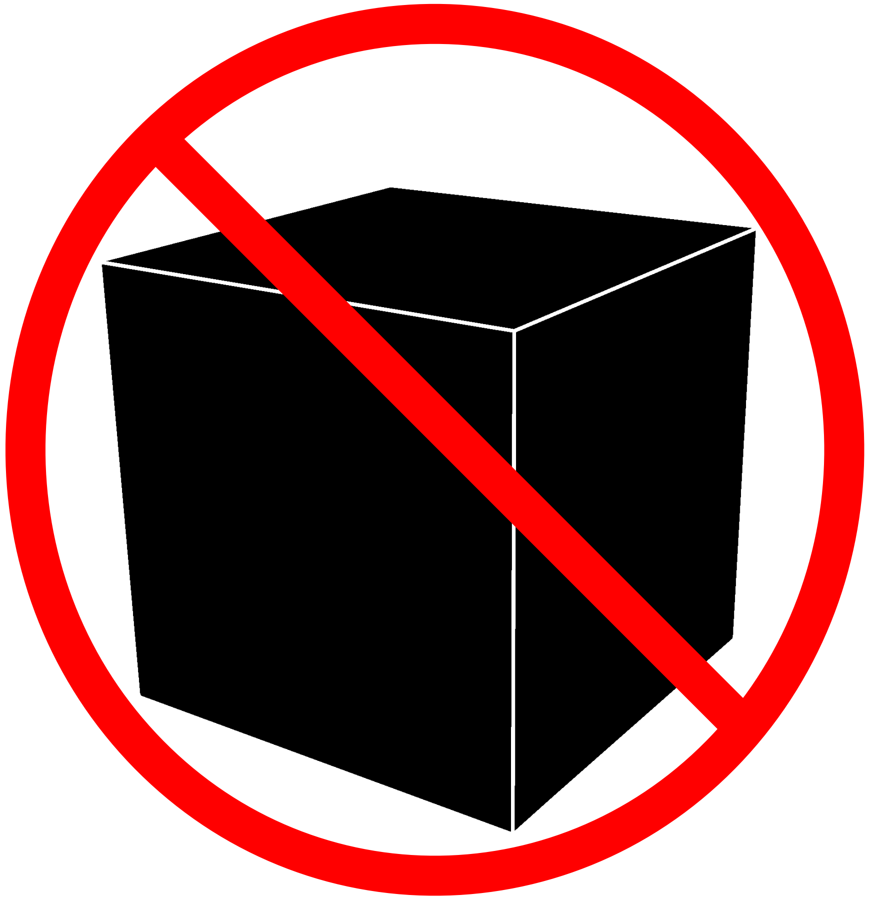
73. What were your usual grades in high school?
□ A □ B ☑ C □ D □ E/F □ Did Not Attend

The Dangers of Black Boxes

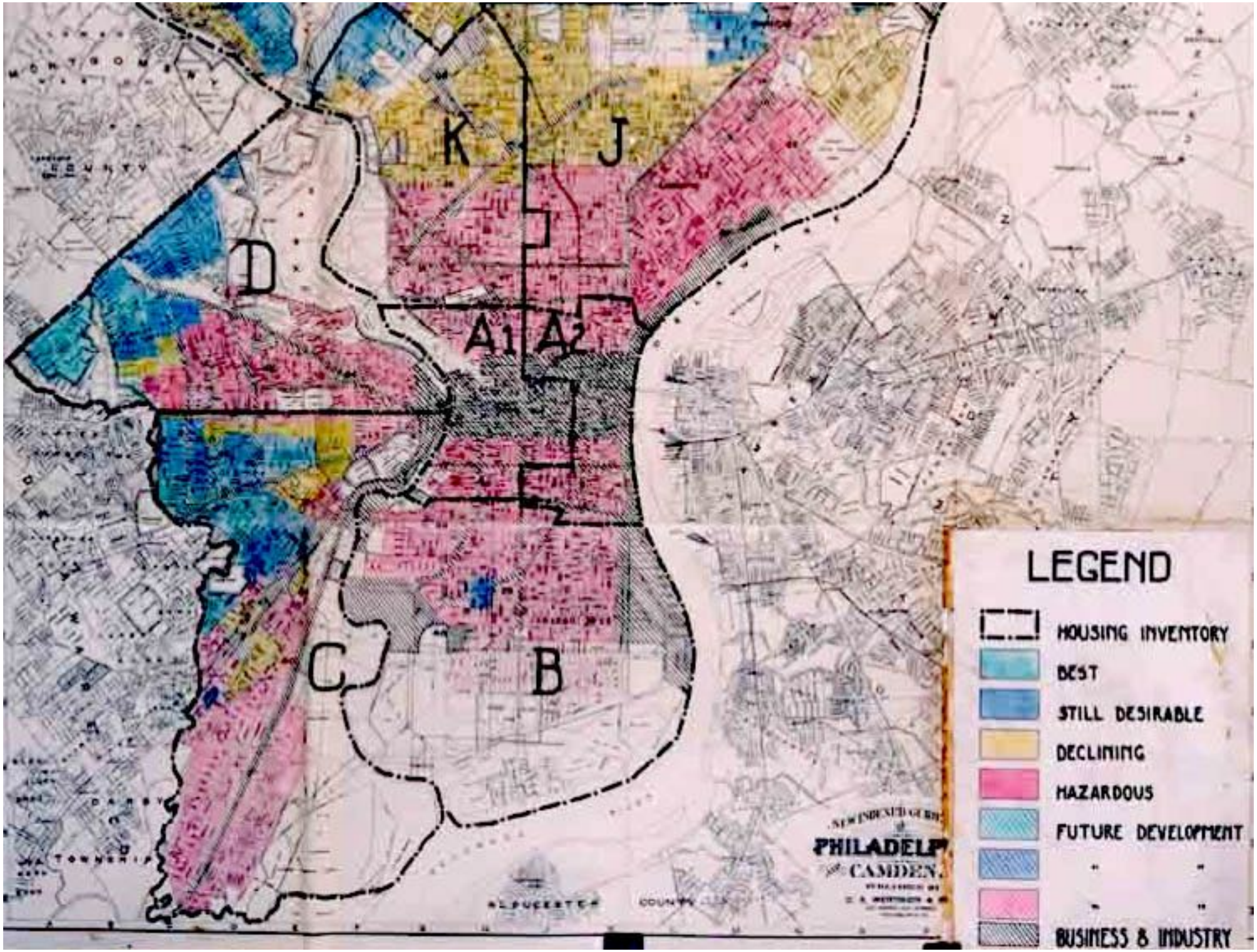
We know what kind of algorithm COMPAS is (not that sophisticated) but we don’t know how much weight it gives to each question, or why

“Environmental” questions (upbringing, family, neighborhood) might be useful for recommending social services, but they should play no role in pretrial, sentencing, or release: your treatment by the system should not depend on things you can’t control

Potential for bias against low-income people, people of color, even though it doesn’t use race directly



Proxies and Redlining



The Dangers of Black Boxes

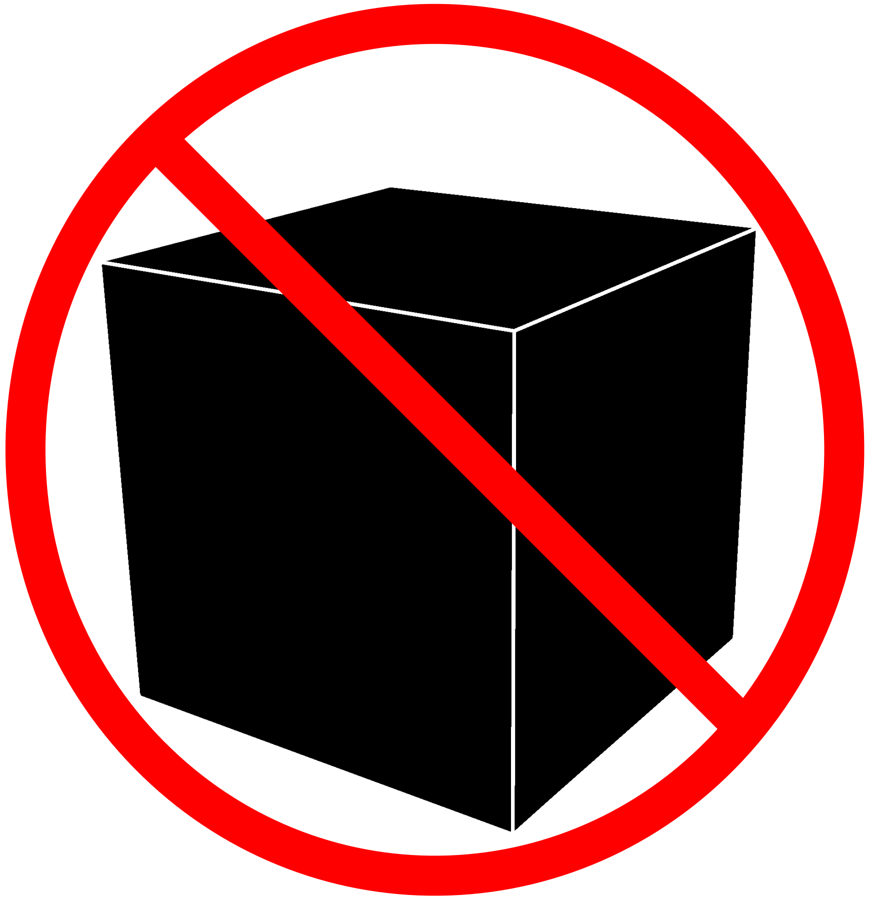
COMPAS produces a “risk score” 1–10, from “low risk” to “high risk”

But we have no way to independently validate its accuracy

COMPAS is expensive to taxpayers

Questionnaire often not completed

Defendants have no explanation of their scores, or what factors contributed: without a license, they can’t even see how their scores depend on the inputs



The Dangers of Black Boxes

Glenn Rodriguez denied parole after COMPAS score of “high risk”

Score was based on incorrect data given to COMPAS by prison staff

Prison staff admitted their mistake, but never updated his score

Since COMPAS is a black box, he was given no explanation

Since he did not have a license to access COMPAS, he was not even able to tell the Parole Board *what his score would have been* if his data had been corrected

Parole board overturned COMPAS’ recommendation two years later

Arnold Public Safety Assessment (PSA)

Specifically for pretrial: gives scores for FTA (Failure to Appear) and NCA (New Criminal Activity, rearrest)

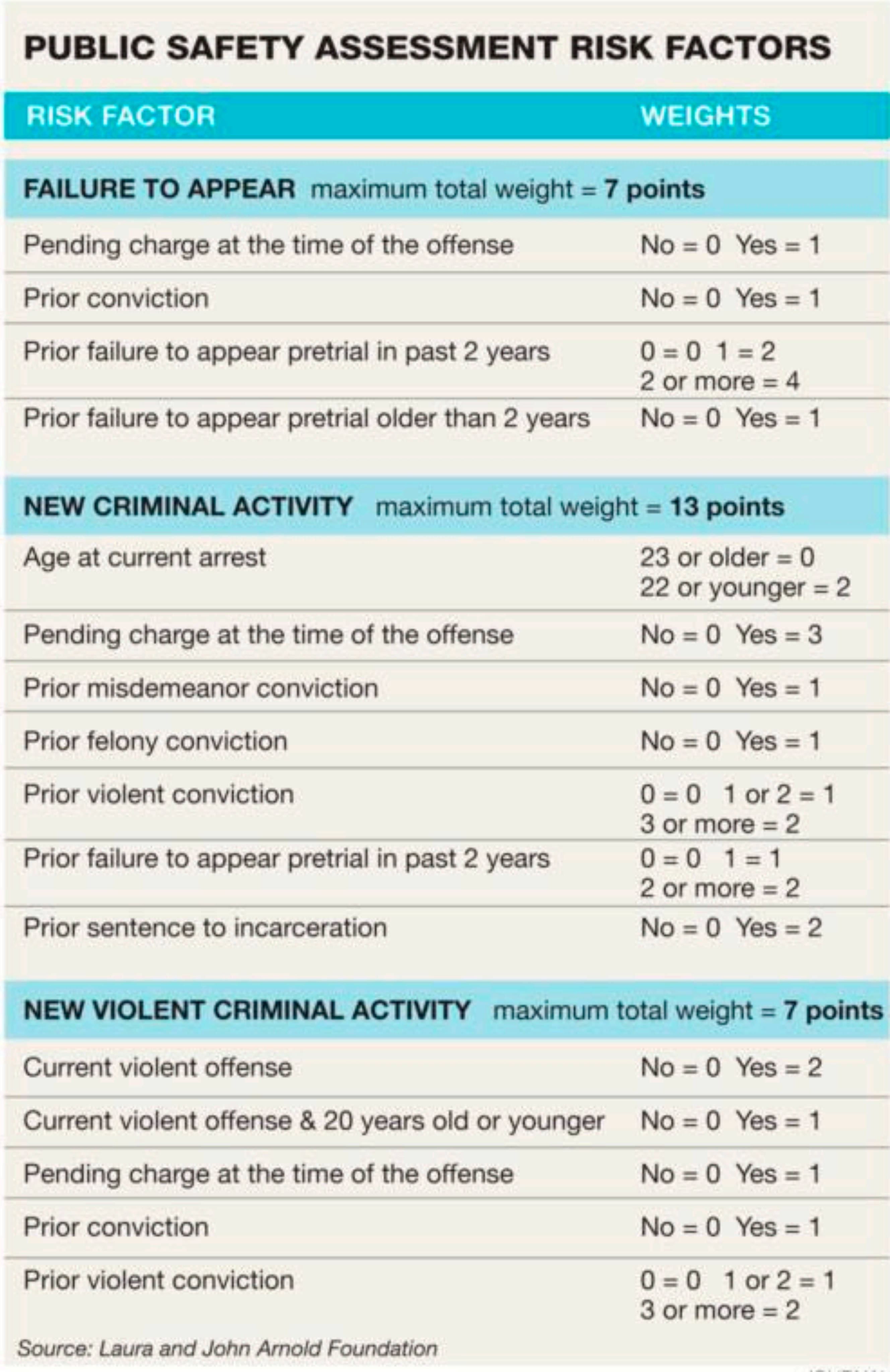
Used in Arizona, Kentucky, Utah, NJ, and about 40 jurisdictions: Bernalillo, Sandoval, San Juan

Not a black box: simple point system, clear explanation of score

No questionnaire, just criminal record: past convictions, past failures to appear

Does not use juvenile record

Uses age but not gender, employment, education, or environment



PUBLIC SAFETY ASSESSMENT RISK FACTORS

RISK FACTOR	WEIGHTS
FAILURE TO APPEAR maximum total weight = **7 points**	
Pending charge at the time of the offense	No = 0 Yes = 1
Prior conviction	No = 0 Yes = 1
Prior failure to appear pretrial in past 2 years	0 = 0 1 = 2 2 or more = 4
Prior failure to appear pretrial older than 2 years	No = 0 Yes = 1
NEW CRIMINAL ACTIVITY maximum total weight = **13 points**	
Age at current arrest	23 or older = 0 22 or younger = 2
Pending charge at the time of the offense	No = 0 Yes = 3
Prior misdemeanor conviction	No = 0 Yes = 1
Prior felony conviction	No = 0 Yes = 1
Prior violent conviction	0 = 0 1 or 2 = 1 3 or more = 2
Prior failure to appear pretrial in past 2 years	0 = 0 1 = 1 2 or more = 2
Prior sentence to incarceration	No = 0 Yes = 2
NEW VIOLENT CRIMINAL ACTIVITY maximum total weight = **7 points**	
Current violent offense	No = 0 Yes = 2
Current violent offense & 20 years old or younger	No = 0 Yes = 1
Pending charge at the time of the offense	No = 0 Yes = 1
Prior conviction	No = 0 Yes = 1
Prior violent conviction	0 = 0 1 or 2 = 1 3 or more = 2

Source: Laura and John Arnold Foundation

Transparency #2: How Well Does it Work in New Mexico?

Local Revalidation

The pretrial services agency should review its risk assessment routinely to verify its validity to the local pretrial defendant population.

“Borrowing” risk assessments from other jurisdictions with no subsequent local validation, basing assessments on subjective stakeholder opinion that is absent research, adopting tools from other criminal justice disciplines for use pretrial, and accepting opaque screening criteria all are fatal—and entirely avoidable—flaws to assessing defendant risk.

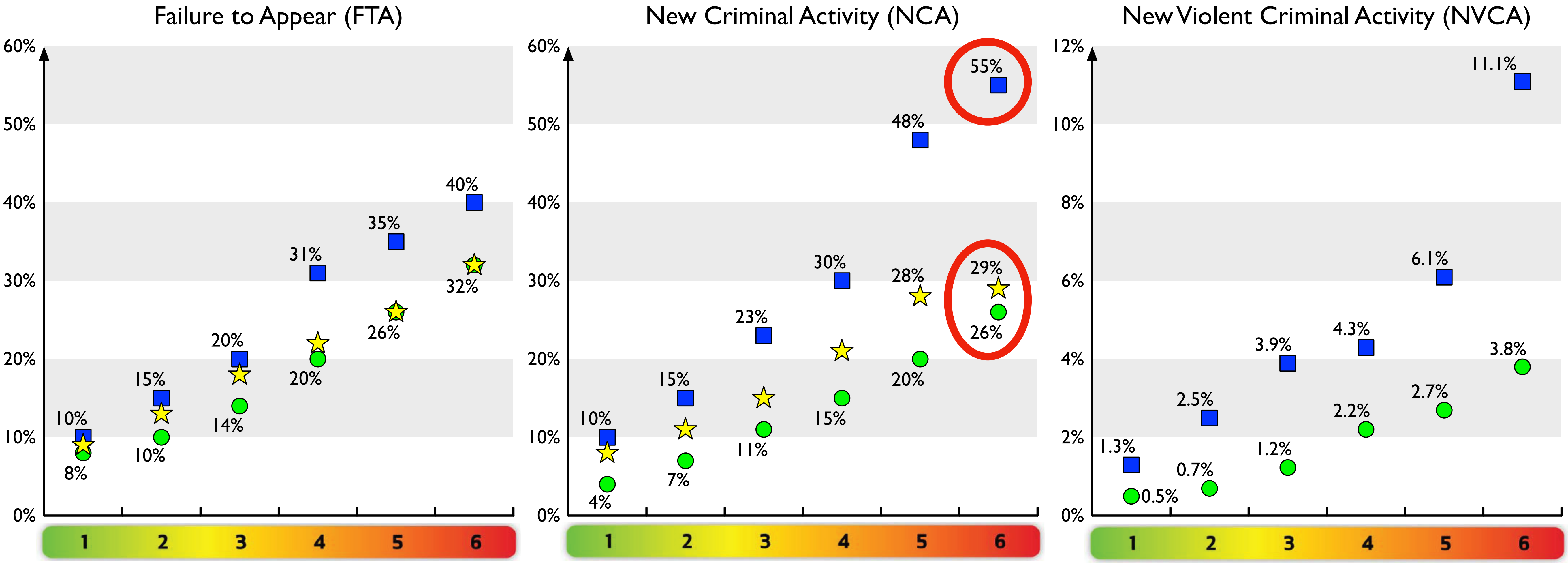
To help ensure race and ethnic neutrality, jurisdictions adopting risk assessments must validate them on the defendant population on which they are used. Validation should gauge the local correlation of race and ethnicity to pretrial failure and risk levels.

National Association of
Pretrial Services Agencies
napsa.org

Local Revalidation

- Every population is different: demographics, implementation…
- Algorithms based on a national data set may perform differently in New Mexico
- Algorithms based on data that is several years old can fail to take the effects of new programs and interventions into account
- Transparency after deployment: does the algorithm perform as expected in New Mexico?
- Validation studies should be done independent of the vendor and the state agency

Comparison between Arnold Foundation's Training Data and Follow-Up Studies in Kentucky and New Mexico



■ Laura and John Arnold Foundation, *Research Summary: Developing a National Model for Pretrial Risk Assessment*

● DiMichele et al., *The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky* (2018)

☆ Ferguson, De La Cerda, and Guerin, *Bernalillo County Public Safety Assessment Review – July 2017 to March 2019*

Policy should be based on risk probabilities, not scores

#3: Detention Should Never Be Algorithmic

Pretrial Supervision Decision Making Framework (Bernalillo County)

FTA: Failure to Appear   NCA: New Criminal Activity

		New Criminal Activity Scale					
		NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
Failure to Appear Scale	FTA 1	(A) ROR	(B) ROR				
	FTA 2	(C) ROR	(D) ROR	(E) ROR-PML 1	(F) ROR-PML 3	(G) ROR-PML 4	
	FTA 3		(H) ROR-PML 1	(I) ROR-PML 2	(J) ROR-PML 3	(K) ROR-PML 4	(L) Detain or Max Conditions
	FTA 4		(M) ROR-PML 1	(N) ROR-PML 2	(O) ROR-PML 3	(P) ROR-PML 4	(Q) Detain or Max Conditions
	FTA 5		(R) ROR-PML 2	(S) ROR-PML 2	(T) ROR-PML 3	(U) Detain or Max Conditions	(V) Detain or Max Conditions
	FTA 6				(W) Detain or Max Conditions	(X) Detain or Max Conditions	(Y) Detain or Max Conditions

United States vs. Salerno (1987)

"In our society, liberty is the norm, and detention prior to trial or without trial is the carefully limited exception"

— Chief Justice Rehnquist

"This case brings before the Court for the first time a statute in which Congress declares that a person innocent of any crime may be jailed indefinitely... if the Government shows to the satisfaction of a judge that the accused is likely to commit crimes... at any time in the future"

— Justice Thurgood Marshall's dissent

New Mexico Constitution, Article II, Section 13, Amended 2016

Bail may be denied by a court of record pending trial for a defendant charged with a felony if the prosecuting authority requests a hearing and proves by clear and convincing evidence that no release conditions will reasonably protect the safety of any other person or the community. An appeal from an order denying bail shall be given preference over all other matters.

A person who is not ~~a danger~~ detainable on grounds of dangerousness nor a flight risk in the absence of bond and is otherwise eligible for bail shall not be detained solely because of financial inability to post a money or property bond.

Individualized Justice

- 1984 Bail Reform Act, U.S. vs. Salerno, and NM Constitution all demand “clear and convincing evidence” of danger to public safety
- An algorithm’s output is not “clear and convincing evidence”
- Algorithms merely summarize information in the criminal record: they don’t provide new information
- Algorithms can only handle *typical* cases, which are similar to many cases in their training data: by definition they cannot handle unusual cases — they are not crystal balls
- Prosecutors can move to detain, and present incriminating evidence: defense attorneys can present exculpatory evidence
- To detain me, you must judge me as an individual, and allow both sides to present evidence about my case

#4: What Does the Data Really Mean?

Beyond Zeroes and Ones

- New Criminal Activity (NCA), Failure to Appear (FTA), and recidivism are often treated as single bits: 0/1, yes/no
- But these fail to tell the full story, or help us understand impact on public safety
  - Failure to Appear: “flight risk” or lack of information, transportation, child care, fear of losing a job…?
  - New Criminal Activity: arrest and crime are not the same thing. Is the new offense major? minor? violent? nonviolent?
  - Recidivism: harm to the public or just a technical violation? (curfew, failure to report, GPS anklets…)
- Validation studies should dig deeper: why did the defendant fail to appear? If they were rearrested, what is the charge?

Feedback Loops

- Computer scientists often view these problems as one-way math problems: predicting behavior from data, ignoring feedbacks
- But this year's predictions affect next year's data. Will this decrease biases over time, or amplify them?
- Predictive policing can reinforce historical patterns, leading to overpolicing in some areas, underpolicing in others
- Need to think about the entire system: humans+algorithms

Prediction vs. Intervention

- The goal is not to *predict* failure, but to help defendants succeed
- Non-technical interventions can help a lot…
  - Text message reminders of court dates, and the consequences of missing them, can reduce Failure To Appear 26–36% [Stanford]
  - Transportation, child care, evening/weekend courts, warrant amnesty courts… help people through the system, de-escalate, and avoid snowballing charges
- In many cases, improvements like these (not “rocket science”) might be just as helpful as a predictive algorithm

Questions?