Bayesian Reasoning Protects Against Confirmatory Biases in Sex Offender Risk Assessment

Richard Wollert, Ph.D., Vancouver, USA
rwwollert@aol.com
http://richardwollert.com
360.737.7712
Any experienced evaluator knows that sex offender (SO) risk assessments elicit a remarkably wide range of opinions from third parties.

- “He’ll reoffend because they always reoffend.”

- “He feels so bad he’ll never reoffend.”
Third parties often share their case-specific theories and interests with risk evaluators.

- Probation officers
- Human services workers
- The defendant’s relatives
- Attorneys
- Other evaluators
A sampler of overt and covert messages received by evaluators.

- “I want you to agree with my assessment.”
- “I’ll be upset if you don’t.”
- “I’ll stop seeing you as a team player.”
- “I’ll stop sending you referrals.”
Evaluators in this context can develop a bias towards confirming a theory they hear about an evaluee before they see him or read his file.

- This bias, in turn, can lead them to collect and interpret information in a way that confirms this presumption.
There is a difference between case-building and confirmatory bias.

- Case-building refers to an intentional process.

- Confirmatory bias “connotes a less explicit, less consciously one-sided case-building process … it refers usually to unwitting selectivity in the acquisition and use of evidence” (Nickerson, 1998, p. 175).
Findley and Scott (2006, pp. 308-309) describe the process of confirmatory bias.

- The foundational tendency is probably best understood as an expectancy bias.
- When people are led to expect some … condition … they tend to perceive that condition in informationally ambiguous situations.
- Effects of having a personal investment in a hypothesis:
  - Confirming information will be overvalued
  - Disconfirming information will be undervalued or missed.
Conditions that amplify the effects of confirmatory bias

- When the source of the hypothesis holds a powerful position.

- When disconfirmation of a favored hypothesis may have a negative effect on the evaluator:
  - A decrease in self-worth; or
  - A decrease professional success.
Confirmatory biases may be unintentional, but they should be controlled in risk assessments for a number of reasons.

- Evaluators testify about their risk estimates as experts in measurement science.
- They must espouse high confidence (a small chance of error) in their estimates.
- The science of measurement proceeds from definition and evidence, not blind faith.
- Confirmatory biases degrade the evidence-based level of confidence that can be placed in estimates.
Bayes’s Theorem (Bayes, 1764), the cornerstone of probabilistic reasoning, is a powerful tool for protecting evaluators against confirmatory biases.

- “It’s nothing personal. It’s just the numbers.”

- BT is also a powerful tool for advancing risk assessment research.
What does Bayes’s Theorem do?

- Estimates the probability a theory is true.
  - Equates to an expert’s level of certainty.

- Estimates the probability a theory is false, (1 minus the probability the theory is true.)
  - Equates to an expert’s level of uncertainty.
What does Bayes’s Theorem do within the context of a risk assessment?

- Two theories are of interest in this context.
  - A. The examinee is a recidivist.
  - B. The examinee is a nonrecidivist.
- BT typically serves 1 of 2 risk assessment goals.
  - Estimate the probability that theory A is true.
  - Estimate the probability that theory B is true.
- This talk focuses on A; the principles apply to B.
What are the terms in Bayes’s Theorem?

- The probability an examinee is disordered (a recidivist) is estimated by combining 2 proportions.
  - The base rate of the disorder (the prevalence of recidivism) in the group from which the examinee has been drawn;
  - The extent to which the evaluator’s risk scale can discriminate between recidivists and nonrecidivists in the examinee’s group.
Evaluators may use one of two types of risk scales.

- Vague and personal (intuitive; clinical judgment)
- More clearly specified and public (“actuarials”)
  - Violence Risk Appraisal Guide (VRAG)
  - Minnesota Sex Offender Screening Tool – Revised (MnSOST-R)
  - Rapid Risk Assessment for Sex Offense Recidivism (RRASOR)
  - Sex Offender Risk Appraisal Guide (SORAG)
  - STATIC-99
  - Automated Sexual Recidivism Scale (ASRS)
Disadvantages of personal assessment systems.

- No standardized definition of risk level.

- Lack of standardization reduces the incentive to collect frequency data for different risk levels.  
  - Net effect: Data are not collected in a reliable way.
Actuarial risk assessment scales, in contrast, specify a range of scores or “risk categories” (high to low).

- An example: Static-99 specifies 4 categories.

- Raw actuarial tables record 2 types of frequency counts from the past.
  - How many recidivists received each score.
  - How many nonrecidivists received each score.
An example of an actuarial table based on frequency counts. The score-wise recidivism rates (first four data rows of the right-hand column) and overall base rate (fifth data row, in bold) are similar to those reported for Static-99.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>Recidivists Only</th>
<th>Nonrecidivists Only</th>
<th>Recidivism Rate for Each Score: A/(A+B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. # Obtaining the Score to the Left</td>
<td>B. # Obtaining the Score to the Left</td>
<td></td>
</tr>
<tr>
<td>3 (H)</td>
<td>36</td>
<td>84</td>
<td>.300</td>
</tr>
<tr>
<td>2 (M_H)</td>
<td>30</td>
<td>140</td>
<td>.176</td>
</tr>
<tr>
<td>1 (M_L)</td>
<td>20</td>
<td>175</td>
<td>.103</td>
</tr>
<tr>
<td>0 (L)</td>
<td>14</td>
<td>301</td>
<td>.044</td>
</tr>
<tr>
<td>All Scores</td>
<td>100</td>
<td>700</td>
<td>.125</td>
</tr>
</tbody>
</table>
Advantages of specified risk assessment scales.

- Data make it possible to estimate base rates (BRs).
  - BR = % of recidivists in a group or subgroup of offenders.
  - BR accuracy requires good sampling, reliable sorting.
- It may also be possible to accurately estimate the discriminative capacity for each test score (next slide).
  - Indicia of these capacities are also called “likelihood ratios” (LR).
- These 2 values may be used as an alternative to frequency counts to calculate the recidivism rates in an actuarial table (Donaldson & Wollert, 2008).
The % of recidivists (R) in each risk category is calculated from counts on slide 17. So is the % of nonrecidivists (R'). R/R' is a score’s LR. The LRs in the last column below are similar to the LRs reported for Static-99.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>Percentage of All Recidivists</th>
<th>Percentage of All Nonrecidivists</th>
<th>Likelihood Ratios: R/R'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R: Who Obtain The Score To The Left</td>
<td>R': Who Obtain The Score To The Left</td>
<td></td>
</tr>
<tr>
<td>3 (H)</td>
<td>.36</td>
<td>.12</td>
<td>3.0</td>
</tr>
<tr>
<td>2 (M_H)</td>
<td>.30</td>
<td>.20</td>
<td>1.5</td>
</tr>
<tr>
<td>1 (M_L)</td>
<td>.20</td>
<td>.25</td>
<td>.8</td>
</tr>
<tr>
<td>0 (L)</td>
<td>.14</td>
<td>.43</td>
<td>.3</td>
</tr>
<tr>
<td>All Scores</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>
Formula for BT. $P(D) = BR \cdot LR^{+}$ = LR for high scores. $P(D|S_h)$ = observed proportion of recidivists among those with high scores. [for more see Waggoner, Wollert, & Cramer (2008, p. 309), Wollert (2007, p. 194), and Wollert & Waggoner (2009, p. 5)].

\[
P(D|S_h) = \frac{P(D)}{1 - P(D)} \times LR^{+}_h \\
1 + \left[ \frac{P(D)}{1 - P(D)} \right] \times LR^{+}_h
\]
The Static-99 actuarial table on slide 17 may be reproduced by combining the BR for the sample with the LRs on slide 19. The following example reproduces the rate for those with high scores.

- **numerator**: \((.125/.875) \times 3.0 = .4286\)

- **denominator**: \(1 + .4286 = 1.4286\)

- **recidivism rate**: \(.4286/1.4286 = 30\%\)
Fagan (1975) determined what the $P(D|S_h)$ values were for each value that might be taken by $P(D)$ and LR.

- He charted the result in a nomogram.

- The advantage of Fagan’s nomogram is that an evaluator may use it to estimate the probability a recidivism theory is correct without calculating BT (Wollert, 2007).

- The next slide shows Fagan’s nomogram (Akobeng, 2006).
Figure 1  The Fagan’s nomogram.
Akobeng described the use of Fagan’s nomogram.

- “A straight line drawn from a patient’s pre-test probability of disease (left axis) through the likelihood ratio of the test (middle axis) will intersect with the post-test probability of disease (right axis).”

- He presented the following example of the use of Fagan’s nomogram.
Figure 2. The use of the Fagan's nomogram (a straight line through the pre-test probability of 10% and the LR+ and the LR+ of 13 yields a post-test probability of 60%).
BRs and LRs have a number of useful applications when they are combined using BT (Beauregard & Mieczkowski, 2009). Here is one.

- Renorm the recidivism rates for one score of an actuarial table using the following method.
  - A. Split (i.e., stratify) a sample into 2 or more groups with different BRs.
    - The stratification method must be reliable.
    - The pattern of LRs must be about the same for all groups.
  - B. Solve BT by combining the BRs for each stratum with the LR for the score.
Model of an unstratified actuarial table that reports the recidivism rate for those with high test scores.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>LR</th>
<th>Recidivism Rate for a Nonstratified (Average) Rate of 17.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>3.0</td>
<td>39%</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>.3</td>
<td>6%</td>
</tr>
</tbody>
</table>
Model of a stratified actuarial that reports rates for SOs with high scores who come from groups with different BRs, but where the average BR is 17.5%.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>LR</th>
<th>Recidivism rate if the group rate is 27%</th>
<th>Recidivism rate if the group rate is 9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>3.0</td>
<td>53%</td>
<td>23%</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>.3</td>
<td></td>
<td>3%</td>
</tr>
</tbody>
</table>
Wollert (2006) applied this method to estimate recidivism rates for sex offenders with high actuarial scores who fell in different age groups.

- The immediate impetus for this study was the finding that sexual recidivism decreases with age (Hanson, 2002).

- This should not have come as a big surprise because criminologists have observed that recidivism declines with age since the early 1800s (Lussier & Healey, 2009).
  - This has been called the “age invariance effect” (Hirschi & Gottfredson, 1983).
The FBI (OJJDP, August 2004) has repeatedly found that the rate of violent offending (including rape) decreases with advancing age.
Hanson (2002) reported the specific effects that advancing age had on long-term sexual recidivism.
In Wollert’s (2006) paper the recidivism rates for those with high scores on various actuarial tests were renormed for 8 of the age groups studied by Hanson:

18-24  25-29  30-34  35-39  40-44  45-49  50-59  60-69

Hanson’s 2002 results reflect an 8-year risk period.

Age-wise base rates were extrapolated from Hanson’s figures and textual information.
A second application is to respecify a set of recidivism rates that have been misspecified for a group of offenders in a stratified test.

- Stratified tests will be misspecified if they are subdivided by an item that is already in the test.

- Low risk offenders will be assigned to high risk groups.
  - Misspecification may not be as apparent when the frequency method is used to calculate recidivism rates.

- A clue to misspecification is that the recidivism rates for scores don’t parallel the recidivism rates for groups.
Hanson (2006) published a stratified table for SOs with high Static-99 scores. The young group has the highest overall rate (row B). It should also have the highest rate for high scorers (row A). It doesn’t.

<table>
<thead>
<tr>
<th>Age Groups</th>
<th>18 to &lt; 25</th>
<th>25 to &lt; 40</th>
<th>40 to &lt; 50</th>
<th>50 to &lt; 60</th>
<th>60 &amp; over</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. 5-year rate for those with high scores</td>
<td>35.5</td>
<td>37.5</td>
<td>25.7</td>
<td>24.3</td>
<td>9.1</td>
</tr>
<tr>
<td>B. 5-year rate for all in the age group</td>
<td>16.2</td>
<td>14.4</td>
<td>8.8</td>
<td>7.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>
This table left readers with the erroneous impression that recidivism rates “peaked” for the 25 to 40 year-olds rather than those 18 to 25.

- Our research team (Waggoner, Wollert, & Cramer, 2008) corrected this error by applying BT to the average LRs for other Static-99 scores and the overall recidivism rate (.162) that Hanson reported for the 18 to 25 year-olds.

- We call the result “Respecified Static-99” (RS-99).
The model for this second application.

<table>
<thead>
<tr>
<th>Test Score</th>
<th>LR</th>
<th>Recidivism rate if the group rate is 16.2%</th>
<th>Recidivism rate if the group rate is 14.4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>3.0</td>
<td>37%</td>
<td>34%</td>
</tr>
<tr>
<td>M</td>
<td>1.1</td>
<td>18%</td>
<td>16%</td>
</tr>
<tr>
<td>L</td>
<td>.3</td>
<td>5.5%</td>
<td>4.8%</td>
</tr>
</tbody>
</table>
The portion of RS-99 that reports the recidivism rates for SOs with high scores is presented below. It shows that recidivism steadily decreases with age. Rates for lower scores are not reported here, but they followed similar trajectories.

<table>
<thead>
<tr>
<th>Age Group</th>
<th>18 to &lt; 25</th>
<th>25 to &lt; 40</th>
<th>40 to &lt; 50</th>
<th>50 to &lt; 60</th>
<th>60 &amp; over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waggoner et al. (2008)</td>
<td>40.3</td>
<td>37.5</td>
<td>25.7</td>
<td>24.3</td>
<td>9.1</td>
</tr>
</tbody>
</table>
Wollert’s estimates for different age groups with high Static scores over 8 years were therefore similar to rates Hanson reported for 5 years. Bayesian analysis is powerful and economical.

<table>
<thead>
<tr>
<th>Age</th>
<th>18 to &lt; 25</th>
<th>25 to &lt; 40</th>
<th>40 to &lt; 50</th>
<th>50 to &lt; 60</th>
<th>60 &amp; over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wollert, 2006</td>
<td>53.6</td>
<td>42.1</td>
<td>32.4</td>
<td>22.8</td>
<td>9.3</td>
</tr>
<tr>
<td>RS-99, 2008</td>
<td>40.3</td>
<td>37.5</td>
<td>25.7</td>
<td>24.3</td>
<td>9.1</td>
</tr>
</tbody>
</table>
A third application is to develop a new actuarial table by combining data from two other tables (call them Tables A and B).

- Estimate one preliminary table by combining the BR of each stratum from Table A with the LRs from Table B.

- Estimate another preliminary table by combining the BR of each stratum from Table B with the LRs from Table A.

- Average the cell-wise rates for the two tables.
  - Requires that both tables be correctly specified.
  - Averaging would be impossible without RS-99.
Our research team recently applied this method within the context of examining the decrease in sexual recidivism that occurs over the life span.

- The original Static-99 risk factor battery addressed the entire issue of age by adding one point to the scores of those who were less than 25 years old at release. Other tests used a similar approach.

- This method of accounting for the age effect is obviously too restricted to be accurate. This has been acknowledged by the developers of Static-99 (Helmus, Thornton, & Hanson, October 2009).
Two stratified tables that consider a wider range of age groups have been disseminated in peer-reviewed journals over the last three years.

- One is “RS-99” (Waggoner et al., 2008).

- The other is the Automated Sexual Recidivism Scale (ASRS; Skelton & Vess, 2008).
Description of the ASRS

- Taps an electronic database of all 5,880 New Zealand SOs released from prison over a 10-year risk period.
- Includes all but 3 Static-99 items.
  - Marital status.
  - Ever victimized non-relatives?
  - Ever victimized strangers?
- Reports recidivism data for six age groups and three score groups (high, medium, and low).
Both RS-99 and ASRS show that recidivism declines with age.

- An advanced understanding of the relationship between age, actuarial scores, and sexual recidivism might be obtained by a two-pronged analysis of these tables.
  - (1) Use BT to combine the data from these tables to develop a new age-stratified table.
  - (2) Contrast the entries in this stratified table with the entries from a table that does not take into account the effect that age has on recidivism.
We (Wollert, Cramer, Waggoner, Skelton, & Vess, Sexual Abuse, in press) compiled one table from the BRs per RS-99 (N=3,425) and the LRs per the ASRS (N=5,880)

- Our other procedures.
  - Compiled a second table from the BRs per the ASRS and the LRs per RS-99.
  - Averaged the cell-wise rates of the two tables to form a new table, the “MATS-1” (N=9,305).
  - Calculated the score-wise rates over all age groups to develop a table that did not account for age effects.
  - Compiled a series of line graphs that compared the results.
Figure 1. MATS-1 recidivism estimates based on age and actuarial scores (L=low, M=medium, H=high) are clearly more accurate than estimates based only on scores.
Conclusions

- The age invariance effect is highly reliable.
- Age-stratified actuarials estimate sexual recidivism more accurately than age-restricted actuarials.
  - The greatest improvement in accuracy is obtained for those with high scores.
- Evaluators should cite data from age-stratified or equivalent actuarial tables in the course of assessing recidivism risk.
Other Comments About Bayesian Reasoning and SO Risk Assessments
Bayesian epistemology. (Bayes’s Theorem) “is a ... way to think about science ... which ... clears up ... misunderstandings ... about the nature of science” (Pigliucci, January 2010, p. 1).

- “Bayes realized that when we think of the probability of a hypothesis to be true we base our judgment on our previous knowledge about the phenomenon under study (i.e., we use induction).
- We then assess new information in light of this prior probability and modify our belief (meant as degree of confidence, not as blind faith) in the hypothesis based on the new information.
- This process can be repeated indefinitely, so that ...
- ... the degree of trust we have in any hypothesis is always due to the current (and ever changing) balance between what we knew before and the new knowledge that additional data bring in.”
Bayesian questions that evaluators might ask themselves as safeguards against the effects of confirmatory biases.

- What is the base rate of recidivism in the population from which this examinee has been drawn?

- What evidence indicates my estimate is accurate? Or is it just “an illusion of certainty” (Wollert, 2007)?
Bayesian questions, continued.

- What specific set of items am I combining with the BR to evaluate the theory the examinee is a recidivist?
- What’s the LR for the examinee’s score on this item set?
- What evidence indicates my LR estimate is accurate?
- Am I relying on the item set with the largest LR as part of my estimation process (Wollert et al., in press)?
Bayesian questions, concluded.

- What risk estimate is suggested by Fagan’s nomogram?
- How can I include probability information in my report?
  - The Appendix (at http://richardwollert.com) includes a short word problem that illustrates the logic of BT.
  - Interested evaluators might consider adopting or modifying some of the passages from the Appendix.
References


Helmus, L., Thornton, D., & Hanson, K. (2009, October). *Should Static-99 recidivism estimates be adjusted based on age at release?* Paper presented at the meeting of the Association for the Treatment of Sexual Abusers, Dallas, TX.


Appendix. The following word problem and its solution illustrates the use of BT.

- A sex offender (SO) incarcerated at the Washington State Reformatory is being evaluated for early release by Dr. Psychologist. Dr. P is using an actuarial called the Risk Tool (RT) to estimate Mr. SO’s recidivism rate.

- Dr. P knows the following facts:
  - 100 of every 1,000 SOs recidivate (i.e., 10%).
  - Mr. SO has a high score on the RT.
  - 40 of every 100 SO recidivists (40%) get high RT scores.
  - 180 of every 900 SO nonrecidivists (20%) get high RT scores.

- What’s the best estimate of Mr. SO’s recidivism risk?
Appendix. The logic of the solution to this problem goes as follows.

- There are 10 chances in 100 that Mr. SO is a recidivist. There are 40 chances in 100 a recidivist will obtain a high RT score.
  - There are 400 chances of drawing a high scoring recidivist.
- There are 90 chances in 100 that Mr. SO is a nonrecidivist. There are 20 chances in 100 a nonrecidivist will obtain a high RT score.
  - There are 1,800 chances of drawing a high scoring nonrecidivist.
- There are 400 + 1,800 = 2,200 chances of drawing a SO with a high score.
- The probability that Mr. SO is a future recidivist is therefore 400 chances divided by 2,200 chances, or 18%. The probability he won’t recidivate is 1-18% = 82%.