TITLE: Phallometric Assessment of Sex Offenders: A Review of the Clinical Evidence and Guidelines

DATE: 1 September 2011

CONTEXT AND POLICY ISSUES:

Deviant sexual preferences have been recognized as an important factor in underlying deviant sexual behaviour.\(^1\) Deviant sexual preferences have also been identified as a predictor of recidivism.\(^2\) Identification of offenders with deviant sexual preferences is clinically important for identifying pathology, conditions under which an offender is likely to reoffend, and determining appropriate individual treatment needs.\(^1,3\) In Canada, 14% of sex offenders have a new charge or conviction within the first five years of release, and within 20 years, 27% had been charged or convicted of another sexual offense, according to a 2004 report.\(^4\)

Penile plethysmography (also known as phallometric assessment or phallometry; these terms are used interchangeably throughout the report) employs a device used to measure sexual arousal by measuring changes in penis circumference or blood volume in response to different auditory or visual stimuli.\(^1,5\) Phallometric assessment is used to identify offenders with deviant sexual preferences, for example those who are sexually attracted to minors, and to determine whether they have responded to treatment.\(^1\)

Among the issues with phallometric assessment is response magnitude, the ability to suppress response to the test, and lack of standardized stimuli.\(^6\) In addition to phallometric assessment, there are other methods for measuring sexual deviance and predicting risk, including actuarial tools such as the Sex Offender Risk Assessment Guide (SORAG) and the Rapid Risk Assessment of Sex Offender Recidivism (RRASOR) tool.\(^7,8\) However, there are concerns regarding the reliability and validity of some of these tools, such as dependence on self-report or use of ambiguous terms.\(^8,9\)

The purpose of this report is to review the literature on diagnostic and risk assessment accuracy of phallometric assessment compared with other tools, to help inform decision making on sex offender assessment methods.

Disclaimer: The Rapid Response Service is an information service for those involved in planning and providing health care in Canada. Rapid responses are based on a limited literature search and are not comprehensive, systematic reviews. The intent is to provide a list of sources and a summary of the best evidence on the topic that CADTH could identify using all reasonable efforts within the time allowed. Rapid responses should be considered along with other types of information and health care considerations. The information included in this response is not intended to replace professional medical advice, nor should it be construed as a recommendation for or against the use of a particular health technology. Readers are also cautioned that a lack of good quality evidence does not necessarily mean a lack of effectiveness particularly in the case of new and emerging health technologies, for which little information can be found, but which may in future prove to be effective. While CADTH has taken care in the preparation of the report to ensure that its contents are accurate, complete and up to date, CADTH does not make any guarantee to that effect. CADTH is not liable for any loss or damages resulting from use of the information in the report.

Copyright: This report contains CADTH copyright material. It may be copied and used for non-commercial purposes, provided that attribution is given to CADTH.

Links: This report may contain links to other information available on the websites of third parties on the Internet. CADTH does not have control over the content of such sites. Use of third party sites is governed by the owners’ own terms and conditions.
RESEARCH QUESTIONS:

1. What is the diagnostic accuracy of phallometric assessment of sex offenders?

2. What is the comparative test accuracy of phallometric assessment versus other tools for risk assessment of sex offenders?

3. What are the evidence-based guidelines and recommendations for risk assessment of sex offenders?

KEY MESSAGE:

Evidence for the diagnostic accuracy of phallometric assessment is inconsistent. There is limited evidence to support the choice of phallometry over other tools for the diagnosis or risk assessment of sex offenders.

METHODS:

Literature Search Strategy

A limited literature search was conducted on key resources including Medline, PsycINFO, PubMed, The Cochrane Library (2011, Issue 7), University of York Centre for Reviews and Dissemination (CRD) databases, ECRI (Health Devices Gold), Canadian and major international health technology agencies, as well as a focused Internet search. No methodological filters were applied to limit retrieval by publication type for questions 1 and 2. A search filter was used to limit retrieval to guidelines for question 3. Where possible, retrieval was limited to the human population. The search was also limited to English language documents published between January 1, 2001 and August 2, 2011.

Selection Criteria and Methods

One reviewer screened the titles and abstracts of the retrieved publications and evaluated the full-text publications for the final article selection, according to selection criteria presented in Table 1.

Table 1: Selection Criteria

<table>
<thead>
<tr>
<th>Population</th>
<th>Adult sex offenders (any offense)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Phallometric Assessment</td>
</tr>
<tr>
<td>Comparator</td>
<td>Any other risk assessment tool (e.g. Static-99, MnSORT, RRASOR, expert diagnosis)</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Diagnostic Accuracy Recidivism prediction Guidelines</td>
</tr>
<tr>
<td>Study Designs</td>
<td>Health technology assessments, systematic reviews, randomized controlled trials (RCTs), non-randomized studies, and evidence-based guidelines.</td>
</tr>
</tbody>
</table>
Exclusion Criteria

Studies were excluded if they did not meet the selection criteria, were duplicate publications or included in a selected systematic review, or were published prior to 2001.

Critical Appraisal of Individual Studies

The quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. No evidence-based guidelines were identified for critical appraisal. Detailed checklist results are not presented. Instead, strengths and limitations of each included study are summarized and described.

SUMMARY OF EVIDENCE:

Quantity of Research Available

The literature search yielded 155 citations. Upon screening titles and abstracts, 132 citations were excluded and 23 potentially relevant articles were retrieved for full-test review. No potentially relevant reports were identified through grey literature searching. Of the 23 potentially relevant reports, 15 did not meet the inclusion criteria. The study selection process is outlined in a PRISMA flowchart (Appendix 1). Eight non-randomized studies were included in this review. No health technology assessments, systematic reviews, RCTs, or evidence-based guidelines were identified.

Summary of Study Characteristics

Details on study characteristics can be found in Appendix 2.

Study setting

Seven included studies were performed by groups in Canada. One study was conducted in the USA. Four studies were conducted in a clinic setting, two were performed in a teaching hospital, one study involved patients incarcerated in a maximum security military prison, and one took place in a maximum security forensic mental health facility.

Patient population

Three studies included subjects with offenses exclusively against children, while five studies included a range of offenses against both children and adults. Two studies defined child as male or female under 16 years old, one as a person under 14 years old, and two as a person under 12 years of age. Two studies used data from the same patient population, collected between 1982 and 1992. One study used data collected between December 1995 and December 1999. The remaining studies did not indicate the data collection periods.

Index tests and reference standards

All included studies involved penile plethysmography. Three used a mercury strain gauge and three used an indium-gallium strain gauge. Two did not specify the equipment used. Phallometric results were compared to Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnosis, Rapid Risk Assessment of Sex Offender Recidivism.
Phallometry for Sex Offenders

(RRASOR) criteria, the Screening Scale of Pedophilic Interest (SSPI), visual reaction time, the Sex Offender Risk Appraisal Guide (SORAG), the Violence Risk Scale – Sexual Offender version (VRS-SO), or expert judgement. A description on assessment tools is provided in Appendix 3.

Outcomes

Five studies reported on diagnostic accuracy, and four studies reported on accuracy of risk assessment (prediction of recidivism). One study reported both outcomes. Two studies measured diagnostic accuracy as the degree of association between the index test and reference standard. One reported the correlation between test results and victim choice. Two studies reported diagnostic accuracy as receiver operator characteristic (ROC) area under the curve (AUC). The ROC curve represents the trade-off between test sensitivity and specificity, and the AUC is an estimate of the probability of a true diagnosis. An AUC of 1.0 (100%) indicates a test can classify a patient with perfect accuracy, while an AUC of 0.5 indicates a test has a 50% chance of correct classification. One study of diagnostic accuracy reported test specificity and sensitivity. Three studies reported accuracy of risk assessment as ROC AUC. One calculated the odds ratios based on logistic regression analysis for predictive factors. The follow-up period to determine re-offense was under 10 years in two studies (mean follow-up of 8.8 years and 6.9 years) and 20 years in two other studies. Recidivism information was based on Canadian Police Information Centre records in two studies and RCMP arrest records in two studies.

Summary of Critical Appraisal

Individual study strengths and limitations are summarized in Appendix 4.

Participants across studies were generally representative of those who would receive the test in practice, however one study was performed in a military facility which may not be representative of a typical test setting. Three studies failed to clearly describe selection criteria.

Three studies did not use the same version of all diagnostic tests on all participants, depending on the year of assessment. Additionally, for these three studies, a newer version of one of the tests used is now available, so diagnostic or predictive accuracy may not be representative of tools used in current practice.

Incomplete data was not well addressed across the included studies. Five studies were unclear or only partially reported uninterpretable and intermediate test results. Three studies were unclear or did not explain withdrawals from the trial.

None of the included studies explicitly stated that index test results were interpreted without knowledge of the results of reference standard or vice versa; thus, increasing the risk of review bias.

Summary of Findings

Detailed study results and author conclusions can be found in Appendix 5.
**Diagnostic accuracy**

Five studies \(^7,9,11-13\) examined the diagnostic accuracy of phallometric assessment. Two studies found no relationship between diagnosis of pedophilia by phallometric assessment and diagnosis using DSM criteria, \(^7,9\) SSPI\(^9\) or expert judgement. \(^7\) One study \(^11\) found that phallometric assessment accurately identified offenders against young boys but not adolescent or adult women. The same study found that phallometric assessment could accurately identify offenders against young girls but in the opposite direction than expected. Two studies \(^12,13\) reported that the probability of correctly discriminating pedophiles from other sex offenders using phallometric assessment was better than chance (86%, P-value not reported\(^12\) and 65% to 69%, P < 0.01\(^13\)).

**Risk assessment**

Four studies \(^7,14-16\) examined the test accuracy of phallometric assessment compared with other tools for risk assessment of sex offenders. Two studies \(^15,16\) found that phallometric assessment partially predicted recidivism risk. One of these \(^16\) found that phallometrically determined Pedophile Assault Index was useful for predicting sexual recidivism but not violent or general recidivism. This study also found that DSM diagnosis, but not Pedophile Index or SSPI, could predict sexual or general recidivism in the opposite direction than expected (DSM diagnosis decreased the likelihood of recidivism). The other study \(^15\) found that phallometric assessment could predict sexual recidivism depending on the stimuli used. VRS-SO could also predict recidivism, while SSPI could not. \(^15\)

One study found that phallometric assessment, \(^14\) but not DSM diagnosis, was predictive of both sexual and violent recidivism. SORAG assessment was also found to be a risk indicator for recidivism. In contrast, one study \(^7\) found phallometric assessment, DSM diagnosis, and expert judgement were unable to predict recidivism better than chance, while RRASOR was a better predictor.

**Guidelines**

No evidence-based guidelines for the diagnosis or risk assessment of sex offenders were identified.

**Limitations**

Eight studies that met the selection criteria were included in the report. Some selected studies involved older data that may not reflect improvements in diagnostic tests and tools available in current practice. A limited number of studies were available for each reference test comparison, and not all available tools available for risk assessment (for example, Static-99, Minnesota Sex Offender Screening Tool) were examined in the identified literature.

The evidence is also limited by the lack of a consensus on the reference standard for comparison of accuracy. Most studies used the offense committed to determine test accuracy, but it is unclear if this is an appropriate method to determine pathology. The test population was primarily men with offenses against children, therefore, the generalizability for other sexual offenders, such as rapists and sexual sadists, is limited. Recidivism data is only available for those who are caught re-offending, and will therefore be an underestimate of true recidivism rates.
Finally, even where the predictive power of the included assessments was statistically better than chance, the probability of a correct diagnosis was typically less than 70%, which may not be robust enough to be clinically useful.

CONCLUSIONS AND IMPLICATIONS FOR DECISION OR POLICY MAKING:

The results for the diagnostic accuracy of phallometric assessment were mixed. Two studies indicated that phallometric assessment could correctly distinguish pedophiles from other offenders, while one found correct assessment was dependent on the choice of victim. No relationship between individuals diagnosed by different methods was observed.

As a tool for risk assessment, phallometric assessment was generally useful for predicting recidivism, with the exception of one study. However, while study authors agree that phallometry is a useful tool for risk assessment, there is insufficient evidence to support its use over other methods.

No evidence-based guidelines for the diagnosis or risk assessment of sex offenders were identified.

PREPARED BY:
Canadian Agency for Drugs and Technologies in Health
Tel: 1-866-898-8439
www.cadth.ca
REFERENCES:


APPENDICES:

APPENDIX 1: Selection of Included Studies

155 citations identified from electronic literature search and screened

132 citations excluded

23 potentially relevant articles retrieved for scrutiny (full text, if available)

0 potentially relevant reports retrieved from other sources (grey literature, hand search)

23 potentially relevant reports

15 reports excluded:
- irrelevant population (1)
- irrelevant intervention (3)
- irrelevant outcomes (11)

8 reports included in review
## APPENDIX 2: Summary of Study Characteristics

<table>
<thead>
<tr>
<th>First Author, Year, Country</th>
<th>Study Design</th>
<th>Patient Characteristics</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Clinical Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostic Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Wilson, 2011, Canada        | Prospective Random patient sample | 130 sexual offenders with offenses against children  
Mean age: 42.3  
Mean follow-up for recidivism: 8.8 years | Penile plethysmography (mercury strain gauge, Pedophile Index) | Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnosis  
Rapid Risk Assessment of Sex Offender Recidivism (RRASOR)  
Expert judgement | Diagnostic accuracy – association between different diagnostic tests (Cramer's V statistic)  
Prediction of recidivism based on RCMP records – ROC area under the curve (AUC) |
| Kingston, 2007, Canada      | Retrospective Cross-section | 206 men convicted of a hands-on sexual assault against an unrelated male or female child (< 16 years old) | Penile plethysmography (Indium-gallium strain gauge, Pedophile Index and/or Pedophile Assault Index) | DSM diagnosis  
DSM and Penile plethysmography  
Screening Scale of Pedophilic Interest (SSPI) | Test clinical accuracy – number of offenders identified as pedophiles, degree of agreement between tests |
| Letourneau, 2002, USA       | Prospective Observational study | 57 sex offenders from a maximum-security military prison  
Mean age: 34.87  
Offense: NR | Penile plethysmography (mercury strain gauge, Association for the Treatment of Sexual Abusers audio stimuli) | Visual reaction time | Test internal consistency – Cronbach's coefficient alpha  
Test clinical accuracy – agreement between results and victim choice (kappa statistic) |
<table>
<thead>
<tr>
<th>First Author, Year, Country</th>
<th>Study Design</th>
<th>Patient Characteristics</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Clinical Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blanchard, 2001, Canada</td>
<td>Retrospective Cross-section</td>
<td>324 adult participants claiming greater sexual attraction to females age ≥ 17 than to any other class of person</td>
<td>Penile plethysmography (Pedophile Index)</td>
<td>NA</td>
<td>Diagnostic accuracy – ROC AUC, test specificity, test sensitivity</td>
</tr>
<tr>
<td>Looman, 2001, Canada</td>
<td>Prospective Case-control</td>
<td>141 rapists (victims ≥16 years old), 49 child molesters (victims &lt; 12 years old)</td>
<td>Penile plethysmography (mercury strain gauge, age/gender stimuli)</td>
<td>Penile plethysmography (mercury strain gauge, child sexual violence stimuli)</td>
<td>Diagnostic accuracy (ability to distinguish rapists from child molesters) – ROC AUC</td>
</tr>
</tbody>
</table>
| Wilson, 2011, Canada        | See above      | 586 men convicted of a contact sexual offense  
Sadist (DSM criteria): Mean age, 28.9 n = 48  
Nonsadist: Mean age, 38.9 n = 534 | Penile plethysmography (Indium-gallium strain gauge, Pedophile Assault Index, Rape Index and/or Adult Assault Index) | Sex Offender Risk Appraisal Guide (SORAG) | Prediction of recidivism based on RCMP arrest records – ROC AUC, Cox regression analysis |
<table>
<thead>
<tr>
<th>First Author, Year, Country</th>
<th>Study Design</th>
<th>Patient Characteristics</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Clinical Outcomes</th>
</tr>
</thead>
</table>
| Canales, 2009, Canada       | Prospective Cross-section | 124 federally incarcerated sex offenders  
Mean age: 36.7  
Offense:  
Rapists (victim age ≥ 14): n = 45  
Child molester (victim < 14 years): n = 37  
Mixed offense (at least one child and one adult victim): n = 28  
Incest offenders: n = 14  
Mean follow-up for recidivism: 6.9 years | Penile plethysmography (Percent full erection) | Violence Risk Scale – Sexual Offender version  
SSPI | Prediction of recidivism based on Correctional Service of Canada and Canadian Police Information Centre records – ROC AUC |
| Moulden, 2009, Canada       | Retrospective Cross-section | 206 men convicted of a hands-on sexual assault against an unrelated male or female child (< 16 years old)  
Follow-up for recidivism: 20 years | Penile plethysmography (Indium-gallium strain gauge, Pedophile Index and/or Pedophile Assault Index) | Diagnostic and Statistical Manual (DSM) diagnosis  
DSM and Penile plethysmography  
Screening Scale of Pedophilic Interest (SSPI) | Prediction of recidivism based on Canadian Police Information Centre reports – logistic regression analysis for predictive factors (Wald criterion) |

DSM = Diagnostic and Statistical Manual of Mental Disorders, SSPI = Screening Scale of Pedophilic Interest,  
SORAG = Sex Offender Risk Appraisal Guide, RCMP = Royal Canadian Mounted Police, ROC = receiver operator characteristic, RRASOR = Rapid Risk Assessment of Sex Offender Recidivism
### APPENDIX 3: Description of Offender Assessment Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penile Plethysmography</td>
<td>“The phallometric method is a psychophysiological technique for assessing erotic interests in male adults and adolescents. In phallometric tests, for gender and age orientation, the individual’s penile blood volume is monitored while he is presented with a standardized set of laboratory stimuli depicting male and female children, pubescents and adults. The patient’s penile blood volume increases (i.e., degrees of penile erection) are taken as an index of his relative attraction to different classes of persons.”¹² p. 118</td>
</tr>
<tr>
<td>Diagnostic and Statistical Manual of Mental Disorders (DSM)</td>
<td>“The Diagnostic and Statistical Manual of Mental Disorders specifies three criteria that must be met to apply a diagnosis of pedophilia. Criterion A requires that the individual has experienced recurrent, intense sexually arousing fantasies, sexual urges, or behaviors involving sexual activity with a prepubescent child or children (generally aged 13 years or younger) over a period of at least 6 months. Criterion B states that the person must have acted on these sexual urges, or the sexual urges or fantasies caused marked distress or interpersonal difficulty. Lastly, Criterion C requires that the individual being assessed is at least 16 years old and at least 5 years older than the child or children in Criterion A.”⁹ p. 423</td>
</tr>
<tr>
<td>Rapid Risk Assessment of Sex Offender Recidivism (RRASOR)</td>
<td>“The RRASOR consists of four easily scored items (age – greater or less than 25, number of previous sexual offenses, unrelated victims, and male victims)”⁷ p. 265</td>
</tr>
<tr>
<td>Screening Scale of Pedophilic Interest (SSPI)</td>
<td>“The SSPI is a brief screening instrument based on historical/static offence variables. The scale includes four items: (1) presence of a male victim; (2) more than one victim; (3) victim is 11 years-old or younger; and (4) unrelated victim.”⁹ p. 427</td>
</tr>
<tr>
<td>Visual Reaction Time (VRT)</td>
<td>“[A] new measure of sexual interest has been developed based on visual reaction time (VRT), the length of time participants view slides presenting clothed models of different ages and gender.”¹¹ p. 208</td>
</tr>
<tr>
<td>Sex Offender Risk Appraisal Guide (SORAG)</td>
<td>“The measure [SORAG] consists of 14 items assessing child and adolescent adjustment, criminal history, psychopathy, and atypical sexual interests: living with both parents until age 16, elementary school maladjustment, history of alcohol problems, marital history, extent of nonviolent offense history, extent of violent offense history, previous sexual offense history, sex and age of index sexual victim, failure on prior conditional release, age at index offense, meeting DSM criteria for any personality disorder, meeting DSM criteria for schizophrenia, phallometrically measured atypical sexual arousal, and Psychopathy Checklist-Revised score.”¹⁴ p. 576</td>
</tr>
<tr>
<td>Violence Risk Scale – Sexual Offender version (VRS - SO)</td>
<td>“The VRS-SO consists of 7 static and 17 dynamic risk predictors empirically, conceptually, or theoretically related to increased risk for sexual recidivism. All items are rated on a 4-point Likert-type scale ranging from 0 to 3.”¹⁵ p. 479</td>
</tr>
</tbody>
</table>
## APPENDIX 4: Summary of Critical Appraisal

<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostic Accuracy</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Wilson, 2011                   | • Patients were representative of those who would receive the test and selection criteria clearly described.  
                                 • All tests were described with sufficient detail to permit replication | • It is unclear whether the index test results were interpreted without the knowledge of results of the reference standard  
                                 • Uninterpretable or intermediate test results were only partially reported |
| Kingston, 2007                 | • Patients were representative of those who would receive the test and selection criteria clearly described.  
                                 • All study withdrawals were explained | • It is unclear whether the index test results were interpreted without the knowledge of results of the reference standard  
                                 • Uninterpretable or intermediate test results were only partially reported  
                                 • Not all patients received the same version of the diagnostic test and more recent versions are available for clinical practice |
| Letourneau, 2002               | • All patients received the same diagnostic tests, and tests were independent of one another  
                                 • Uninterpretable or intermediate test results were reported and any withdrawals were explained  
                                 • Tests were described in sufficient detail to permit replication | • Patients were recruited from a military facility and may not represent the general population; selection criteria were not clearly described  
                                 • It is unclear whether the index test results were interpreted without the knowledge of results of the reference standard |
| Blanchard, 2001                | • Patients were representative of those who would receive the test and selection criteria clearly described  
                                 • All patients received the same diagnostic test, and test was described in sufficient detail to permit replication | • No reference standard was used. Accuracy was based on “credible accusation”  
                                 • It is unclear whether index test results were interpreted without knowledge of the accusations  
                                 • Uninterpretable or intermediate test results were not clearly reported |
| Looman, 2001                   | • Selection criteria was clearly described  
                                 • All patients received the same diagnostic tests, and tests were described in sufficient detail to permit replication  
                                 • Uninterpretable or intermediate test results were reported | • Exclusion of some participants may result in a population not representative of those who would receive the test in practice  
                                 • It is unclear whether the index test results were interpreted without the knowledge of results of the reference standard |
| **Risk Assessment**            |           |             |
| Wilson, 2011                   | See above |             |
| Kingston, 2010                 | • Patients were representative of those who would receive the test  
                                 • All tests were described in sufficient detail to permit replication | • Not all patients received the same version of the diagnostic tests, and more recent versions are now available for clinical practice  
                                 • It is unclear whether the index test results |
<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Patients were representative of those who would receive the test</td>
<td>were interpreted without the knowledge of results of the reference standard</td>
</tr>
<tr>
<td></td>
<td>• All patients received the same diagnostic tests, and tests were</td>
<td>• Comparator tests were not independent of one another</td>
</tr>
<tr>
<td></td>
<td>described in sufficient detail to permit replication</td>
<td>• Uninterpretable results were partially reported</td>
</tr>
<tr>
<td>Canales, 15 2009</td>
<td>• Comparator tests were independent of one another</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Selection criteria was not clearly described</td>
<td>• Uninterpretable or intermediate test results were not clearly reported</td>
</tr>
<tr>
<td></td>
<td>• Not all patients received the same version of the diagnostic tests, and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>more recent versions are now available for clinical practice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• It is unclear whether the index test results were interpreted without</td>
<td></td>
</tr>
<tr>
<td></td>
<td>the knowledge of results of the reference standard</td>
<td></td>
</tr>
<tr>
<td>Moulden, 16 2009</td>
<td>• Patients were representative of those who would receive the test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• and selection criteria clearly described</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Comparator tests were independent of one another, and described in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sufficient detail to permit replication</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX 5: Study Results and Author Conclusions

<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Main Study Findings</th>
<th>Authors’ Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostic Accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilson, 2011</td>
<td>Association between diagnostic measures (Cramer’s V statistic): Penile plethysmography (PPG) and Rapid Risk Assessment of Sex Offender Recidivism (RRASOR): 0.10, P = NS PPG and Diagnostic and Statistical Manual of Mental Disorders (DSM): 0.04, P = NS PPG and expert judgement: 0.19, P = NS DSM and RRASOR: 0.19, P = NS DSM and expert judgement: 0.49, P &lt; 0.001 RRASOR and expert judgement: 0.31, P &lt; 0.003</td>
<td>“It was found that expert ratings, DSM-IV-TR diagnosis, and phallometric testing were unrelated to each other and unrelated to recidivism. In regression analysis, the inclusion of these items did not add to the prediction of recidivism when the RRASOR scores were already accounted for.” p. 271</td>
</tr>
<tr>
<td>Kingston, 2007</td>
<td>PPG (Pedophile Index [PI] or Pedophile Assault Index [PAI] score ≥ 1) identified 110 offenders as pedophiles and 45 offenders as non-pedophiles. DSM criteria identified 85 offenders as pedophiles and 79 offenders as non-pedophiles. DSM + PPG identified 49 offenders as pedophiles and 43 offenders as non-pedophiles. Screening Scale of Pedophilic Interest (SSPI) identified 103 offenders as pedophiles and 103 offenders as non-pedophiles. PI was statistically significantly associated with pedophiles diagnosed according to the DSM (P = 0.021). 24.6% of non-pedophiles and 91% of pedophiles were predicted, for an overall success rate of 61.9%. For every unit increase in the PI score, the predicted odds of a diagnosis of pedophilia by DSM criteria increased 32%. PI score was not significantly associated with diagnosis by</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“[P]rocedures used to define pedophilia were not significantly related to one another.” p. 423</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Given the importance of deviant arousal in managing sexual offenders for child molesters in particular, emphasis should be placed on this indicator of preference. Until there is evidence that the diagnosis of pedophilia, according to DSM criteria, offers some demonstrable utility regarding such aspects as group differentiation and recidivism, various treatment and management strategies should be guided by purely behavioural and/or physiological indicators of preference.” p. 434</td>
<td></td>
</tr>
<tr>
<td>First Author, Publication Year</td>
<td>Main Study Findings</td>
<td>Authors’ Conclusions</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Letourneau, 11 2002</td>
<td>SSPI. (PPG was used as diagnostic criteria for other methods and was therefore not included in logistic regression models for these methods)</td>
<td>“Results indicated adequate internal consistency for both measures.” p. 207</td>
</tr>
<tr>
<td></td>
<td>There was no significant relationship between individuals diagnosed by each of the methods examined in the study.</td>
<td>“Convergent validity and an assessment of clinical usefulness for both measures indicated that (a) both measures accurately identified offenders against young boys; (b) the VRT, but not the PPG, significantly identified offenders against adolescent girls; (c) neither measure reached statistical significance in identifying offenders against adult women; and (d) the VRT did not reach statistical significance in identifying offenders against young girls and the PPG did reach statistical significance but in the opposite direction as was expected.” p. 207</td>
</tr>
</tbody>
</table>

Letourneau, 11 2002

Internal consistency:
Cronbach’s coefficient alpha (< 0.70 unacceptably low)
PPG: >0.70 for all stimulus categories
Visual reaction time (VRT): >0.70 for all stimuli except males aged 2 to 4

Convergent validity (correlation between results and victim choice)
PPG, stimuli
Female child: -0.281, P < 0.05
Male child: 0.653, P < 0.01
Female adolescent: 0.184, P = NS
Female adult: 0.212, P = NS

VRT, stimuli
Female child: 0.082, P = NS
Male child: 0.687, P < 0.01
Female adolescent: 0.281, P < 0.05
Female adult: 0.012, P = NS

Clinical accuracy (Kappa statistic):
PPG, offense against:
Girls aged 0 to 13: -0.369, P < 0.01
Boys aged 0 to 13: 0.614, P < 0.001
Girls aged 14 to 17: 0.192, P = NS
Women aged 17+: 0.208, P = NS

VRT, offense against:
<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Main Study Findings</th>
<th>Authors’ Conclusions</th>
</tr>
</thead>
</table>
| **Blanchard,** 12 2001       | Diagnostic Accuracy (ROC AUC)  
AUC 0.8619; 95% CI, P-value not reported  
Test specificity: 96%  
Test sensitivity: 61% (minimum) | “Specificity, calculated for those who had sexual contact with the most women and thus the most evidence of attraction to them was 96%. Sensitivity, calculated analogously for men with the most offenses against children, was 61%.” p. 118 |
| **Looman,** 13 2001          | Diagnostic Accuracy (ROC AUC)  
Age/gender Penile plethysmography:  
AUC 0.65, 95% CI not provided, P < 0.01  
Child sexual violence assessment:  
AUC 0.69, 95% CI not reported, P < 0.001  
The difference in predictive accuracy between the two assessment methods was not statistically significant (P-value not reported) | “Analyses revealed that the audiotaped child sexual violence assessment was as effective as the slide [age/gender] assessment in discriminating child molesters from rapists. It was also found that rapists displayed some degree of deviant response to both stimulus sets.” p. 3 |
| **Risk Assessment**           | Predictive Accuracy (ROC AUC)  
PPG:  
AUC 0.41, 95% CI 0.235 to 0.579, P = NS  
DSM:  
AUC 0.54, 95% CI 0.373 to 0.704, P = NS  
RRASOR:  
AUC 0.67, 95% CI 0.498, 95% CI 0.839, P < 0.05  
Expert judgement:  
AUC 0.55, 95% CI 0.389 to 0.713, P = NS | “It was found that expert ratings, DSM-IV-TR diagnosis, and phallometric testing were unrelated to each other and unrelated to recidivism. In regression analysis, the inclusion of these items did not add to the prediction of recidivism when the RRASOR scores were already accounted for.” p. 271 |
| **Kingston,** 14 2010         | Predictive Accuracy (ROC AUC)  
Sexual recidivism:  
DSM: AUC 0.54, 95% CI 0.47 to 0.62, P = NS  
PPG: AUC 0.60, 95% CI 0.53 to 0.67, P < 0.01 | “The results support the use of more behaviourally operationalized indicators of sexual sadism, especially phallometric assessment of sexual arousal, and suggest the DSM criteria for sexual sadism require further work” p. 574 |
<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Main Study Findings</th>
<th>Authors’ Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canales, 2009</td>
<td>Predictive Accuracy (ROC AUC) of sexual recidivism (conviction)</td>
<td><strong>“The VRS-SO Sexual Deviance factor and some phallometric indexes of deviant arousal (e.g., female children) were predictive of sexual recidivism whereas the SSPI was not. The pattern of findings across analyses broadly supports the construct validity of the VRS-SO in assessing sexual deviance.”</strong> p. 474-5</td>
</tr>
</tbody>
</table>
|                              | Aggregate sample (n = 124) | PPG adult female stimuli: 0.54, 95% CI 0.43 to 0.65  
PPG child female stimuli: 0.62, 95% CI 0.51 to 0.74  
PPG pubescent female stimuli: 0.63, 95% CI 0.52 to 0.73  
PPG child male stimuli: 0.56, 95% CI 0.45 to 0.68  
PPG pubescent male stimuli: 0.53, 95% CI 0.41 to 0.65  
Violence Risk Scale-Sexual Offender version (VRS-SO): 0.60, 95% CI 0.49 to 0.71  
SSPI: NR | |
|                              | Child victims only (n = 79) | PPG adult female stimuli: 0.58, 95% CI 0.43 to 0.72  
PPG child female stimuli: 0.65, 95% CI 0.50 to 0.81 | |
<table>
<thead>
<tr>
<th>First Author, Publication Year</th>
<th>Main Study Findings</th>
<th>Authors’ Conclusions</th>
</tr>
</thead>
</table>
| Moulden, 2009                 | PPG pubescent female stimuli: 0.66, 95% CI 0.38 to 0.69  
PPG child male stimuli: 0.68, 95% CI 0.54 to 0.82  
PPG pubescent male stimuli: 0.68, 95% CI 0.53 to 0.82  
VRS-SO: 0.67, 95% CI 0.54 to 0.80  
SSPI: 0.48, 95% CI 0.33 to 0.62 | "The results suggest that although phallometry may distinguish recidivists and non-recidivists, those individuals determined to be pedophiles, regardless of definition, do not recidivate more often or more quickly than non-pedophiles." p. 698 |

63% of sexual recidivists and 96% of non-recidivists were correctly classified (overall success rate 87%). DSM and PAI predicted sexual recidivism; however DSM diagnosis decreased the likelihood of recidivism.

DSM: Odds Ratio (OR) 0.33, 95% Confidence Interval (CI) 0.12 to 0.88  
PI: OR 0.95, 95% CI 0.69 to 1.32  
PAI: OR 2.06, 95% CI 1.06 to 4.00  
SSPI: OR 0.88, 95% CI 0.47 to 1.65  

Different diagnosis methods did not predict violent recidivism

89% of general recidivists and 94% of non-recidivists were correctly classified (overall success rate of 91%)

DSM: OR 0.11, 95% CI 0.03 to 0.47  
PI: OR 0.99, 95% CI 0.66 to 1.48  
PAI: OR 1.46, 95% CI 0.61 to 3.50  
SSPI: OR 1.06, 95% CI 0.46 to 2.48

AUC = area under the curve, CI = confidence interval, DSM = Diagnostic and Statistical Manual of Mental Disorders, HR = hazard rate ratio, NS = not significant, OR = Odds Ratio, PAI = Pedophile Assault Index, PI = Pedophile Index, PPG = penile plethysmography, ROC = Receiver Operator Characteristic, RRASOR = Rapid Risk Assessment of Sex Offender Recidivism, SORAG = Sex Offender Risk Appraisal Guide, SSPI = Screening Scale of Pedophilic Interest, VRT = visual reaction time