THE STRONG-R PILOT ASSESSMENT STUDY

Washington State Department of Corrections

April 26, 2016

Completed by

Zachary Hamilton, Ph.D.,

Alex Kigerl, Ph.D.,

&

Douglas Routh, MA

The Washington State University Department of Criminal Justice

&

Criminology, Institute for Criminal Justice

EXECUTIVE SUMMARY

As part of the implementation process of the STRONG-R, knowledge and assurances were desired by the Washington State Department of Corrections (WADOC) regarding the consistency of item scoring and the application of risk category cut points. While the STRONG-R was initially developed using an extremely large and robust sample (N~45,000) there was a concern that the implementation of additional training and the inclusion of offender needs assessment data as part of risk assessment scoring may result in offender scores and risk category proportions that are inconsistent with estimates provided via the initial development sample. To provide estimates of future scoring and risk categories proportions a sample of current offenders was assessed as part of a STRONG-R pilot. The current report provides findings of the STRONG-R Pilot Assessment Study.

As part of the study, a group of 45 WADOC staff was trained to assess offenders using the new STRONG-R tool. A random sample of 350 offenders was drawn from a January 2016 snapshot of current WADOC offenders in both prison and community settings. A total of 200 offenders identified within this draw were available to be assessed for the pilot study. Pilot data were recorded by staff and an algorithm designed by the WSICJ was created to convert assessment responses into a workable dataset. Data were then coded and analyzed. To assist the analysis, additional development data was gathered, adding offenders assessed since the initial development sample was collected. Additional data gathered provided needed information on incarcerated offenders and increased the stability of recidivism estimates. The updated development sample consists of nearly 100,000 offenders assessed by the WADOC between August of 2008 and October of 2015. Two-year recidivism assessments were also provided for all offenders that have been assessed and have reentered the community between August of 2008 and October of 2013. A survey was also distributed and analyzed to provide stakeholder insight with regard to future STRONG-R implementation efforts.

Study analyses provided descriptive comparisons of development and pilot sample on current STRONG-R item responses. These item level comparisons demonstrated consistency of content across the two samples. However, specific items relating to correctional events and items of lower prevalence were identified to possess some instances of inconsistent scoring. It is recommended that these items be reexamined by stakeholders and may be outlined for removal or modification.

Next, continuous risk scores were computed for males and females on four recidivism outcomes – violent, property, drug, and 'any' felony conviction. Three cut point options were established using the development sample. Risk category proportions are provided for the following STRONG-R categories – 1) Lower, 2) Moderate, 3) High Drug, 4) High Property, 5) High Violent, and 6) High Violent, Property, and Drug (HVPD)[1]. To provide a reference to current labor and contact standards category proportions for the Static Risk Assessment – Version 2 (SRA2) were also provided. These risk scores and category cut points were then applied to the pilot data and additional comparisons were made regarding recidivism estimates and labor impact.

[1] Readers should note that the WADOC altered the original name of "Criminally Diverse" to "High Violent, Property, and Drug" (HVPD).

Findings revealed the following:

- STRONG-R risk categories can be established that provide an intrinsic meaning with regard to recidivism and also provide similar category proportions to that of the SRA2.
- Identified in the report as Option 3, the WSICJ established a cut point for high violent, property and drug models that identifies offenders with a recidivism probability that is twice that of the base rate of the given model's conviction estimate. The lower risk offender category is set at one-fifth the recidivism base rate for 'any' felony conviction.
- Using the outlined STRONG-R Option 3 for both male and female offenders, category proportions are similar to that of the SRA2, relieving concerns of increased labor demands as a result of implementation. Furthermore, the cut point system results in category proportion's that reduce bias, preventing the over classification of female offenders into high risk categories.
- When examining risk category performance, it was found that, compared to the SRA2, the STRONG-R risk categories provide greater accuracy via improved discrimination. That is, even though the category proportions are similar to that of the SRA2, fewer Lower and Moderate risk offenders were identified to recidivate and a greater proportion of high risk offenders were found to recidivate. Thus, the STRONG-R risk categories will provide a more accurate assessment of future offending prediction and a more efficient use of WADOC supervision labor as a result.
- The new STRONG-R high violent, property, and drug risk category - HVPD - provides an indicator of offenders that are predicted to be 'opportunistic' in their future criminal activity. This category is a novel element of the STRONG-R system that should prove to improve case management and planning practices.
- Comparing the SRA2 to the STRONG-R risk categories by race/ethnic groupings revealed reductions in disproportionality of risk classifications across race/ethnic categories; however, substantial variation still remains. That said, accuracy of the risk models remained consistent, finding only one instance (in 20) of a significant Area Under the Curve (AUC) variation from the 'overall' aggregated development sample.
- Survey findings by assessors generally revealed confidence in the tool's use and the duration of time needed to complete the interview and scoring. Several notable items were identified to be addressed for training prior to implementation

Overall the STRONG-R pilot identified several positive signals of an accurate and efficient assessment tool. While many notable areas of improvement were identified, this important feedback can be used to improve training and sustainability going forward. Nest steps are identified with regard to the pilot study and remaining contractual deliverables. Further examination of difficult and potentially unreliable items is recommended. Next, a study examining the tool's inter-rater reliability will be completed along with the creation of training and quality assurance procedures. Finally, the WSICJ is currently examining the relationship between the STRONG-R needs assessment and potential linkages, sequencing, and service gaps for WADOC provided interventions and treatments.

Introduction

Following the development of the Static Risk and Offender Needs Guide – Revised (STRONG-R) risk and needs assessment, preparations were needed to assure proper implementation fidelity and reliability of the tool in practice. On January of 2016, the Washington State Department of Corrections (WADOC) partnered with Washington State University's (WSU) Institute for Criminal Justice (WSICJ) to craft methods and analyze: 1) a piloting of the STRONG-R risk assessment, 2) assess interrater reliability, 3) provide quality assurance procedures, and 4) identify a menu of interventions to be used in conjunction with the STRONG-R needs assessment. The current report provides a description and findings form the first project deliverable – the STRONG-R pilot study.

Project Deliverable Description

With the training and implementation of the STRONG-R scheduled to be completed within the next year, there is a potential to create substantial changes among risk categories that will impact general day-to-day operations. Specifically, while estimates have been created using the STRONG-R development samples, a potential exists in which newly scored offenders may be more likely to be scored and categorized into a higher (or lower) risk category. Currently, the Static Risk Assessment – Version 2 (SRA2) categorizes offenders into four categories – Lower, Moderate, High Non-Violent (HNV), and High Violent (HV). The STRONG-R is designed to implement a similar, yet slightly more specified categorization including: Lower, Moderate, High Drug, High Property, High Violent, and High Violent-Property-Drug (HVPD). Given that the WADOC adjusts supervision and intervention strategies based on risk level, alterations to the proportions of offenders in each risk category has the potential unintended consequence of altering labor needs, especially for those offenders in the community.

Pilot Sample

To account for potential modifications, a pilot assessment and data analysis of test cases was needed. Working with the WADOC, WSICJ devised a method of pilot testing the STRONG-R. First, a team of assessors was assembled by the WADOC to be trained and assess a relatively small sample of offenders. The team was comprised of 45 assessors with a variety of assessment experience levels and represented offender

change, community and prison locations/divisions. Based on prior assessment administrations it was estimated that the team could perform roughly 250-350 assessments in the project time allotted and this estimate was used to provide a range of the subjects to be selected for the pilot.

Next, a simple random sample was created. The creation of the sample comprised several steps. First, a target population was established along with important sample characteristics to consider. It was determined that the WADOC would draw a *snapshot* of offenders currently supervised in both prison facilities and the community. All current offenders were considered eligible for selection with two exceptions, those classified as *out-of-state* or those serving life sentences (without the possibility of parole). This group of offenders formed our target population.

Subject Matter Experts (SMEs) from the WADOC were then assembled, comprising staff from the community, prison, budget, research, and Advance Corrections. The team was asked to outline a set of key characteristics to consider when evaluating the quality and representation of the subjects selected for the pilot assessment. The group suggested that the sample consider the prevalence of WADOC offenders within the snapshot relating to the following: prison versus community, race/ethnicity, gender, age, SRA2 risk level category, current sex offender status, prison custody level, community corrections region, eligibility for alternative sentences, earned release date, expected community supervision duration, and where available Static 99 and/or Stable sex offender risk categorizations.

On January 21 the snapshot dataset was drawn and provided to WSU researchers, comprising 31,585 eligible offenders. WSU then cleaned and organized the identified offender characteristics into metrics useable for analysis purposes. Descriptive statistics were then assessed for the overall snapshot population as well as a breakdown of prison and community sub-samples.

Following the examination of sample descriptives, areas of over and under-coverage were examined. It was determined that current WADOC policies concerning community-based sex offenders and those eligible for sentencing alternatives may be of concern. A decision was made to *oversample* these two community populations to ensure pilot assessments would return adequate descriptive information of offender's potential new classification categorization. A total of 50 offenders from each of these two unique

subsamples (25 each) were drawn as part of the over sampling procedure. In addition, there was a potential for attrition, where randomly selected offenders perceived to be eligible were later found to not be appropriate. To ensure that the final sample assessed was adequate sample of offenders 25 offenders form both the prison and community samples were selected as *alternates*, to be randomly drawn from in substitute, if it was deemed an originally selected offender could not be utilized or assessed. After the initial return of assessments was examined, it was found that fewer than anticipated female offenders were assessed. Therefore, an additional sample of 34 female offenders was also collected. Finally, it was found that inmates were slightly more prevalent (52%) than community-based offenders (48%) in the snapshot drawn. The random sample was stratified on this key characteristic to ensure similar proportionality among the pilot subjects.

The random samples were then drawn from the prison and community-based populations. A total of 156 inmates were selected, which included the 25 alternates previously discussed. A total of 194 community-based offenders were selected, which also included 25 alternates and the 50 oversampled subjects. Thus, the final sample was outlined to potentially consist of 350 subjects. Given that 50 of these subjects will serve as alternates, the likely total of offenders assessed is anticipated to be 300, which is within the feasible range outlined by the WADOC.

The random samples were then examined for accuracy and statistical anomalies. Characteristics of the random samples were compared to the original and the combined population from which they were drawn. Examining each measure and category revealed that, with the exception of the oversampled sex offender and alternative sentence populations, the random samples differed from the snapshot populations by roughly 0 to 5%. These variations comprise small-to-negligible effect size differences and therefore suggest that the sampling procedures were successful in selecting pilot study subjects that were representative of the current WADOC population on all pre-identified characteristics. Table 1 provides the descriptive statistics of the snapshot populations and the prison and community samples drawn. On January 25th, the list of selected offenders was made available to the WADOC to organize future logistics of pilot assessment.

Table 1. STRONG-R Snapshot and Simple Random Pilot Sample - %/Mean(SE)

Sample frame characteristic	Prison snapshot population (n=169,419)	Prison random sample (n=156)†	Community snapshot population (n=15,166)	Community random sample (n=194)	Combined snapshot population (N=184,585)
Race/Ethnicity					
White	60.6	57.7	69.2	67.0	64.6
Black	17.5	21.2	12.9	14.9	15.5
Hispanic	13.1	14.7	8.6	7.7	10.9
Other	8.8	6.4	9.3	10.3	9.0
Male	92.2	92.9	85.7	90.2	89.2
Age					
60+	4.9	3.8	3.9	5.7	4.7
50-59	12.9	10.3	11.3	16.0	12.6
40-49	21.6	23.7	19.6	22.2	20.8
30-39	33.8	33.3	32.2	30.4	32.7
20-29	26.1	28.8	31.7	24.2	28.3
19-13	0.7	0.0	1.3	1.5	1.0
SRA2					
High violent	52.0	55.1	47.9	42.8	49.9
High non-violent	14.5	12.8	27.1	22.7	20.3
Moderate	11.3	14.1	11.7	16.5	11.5
Low	20.7	17.3	12.6	8.0	17.3
ERD					
6-12 months	22.9	23.2	--	--	22.9
1-2 years	25.9	23.2	--	--	25.9
2-4 years	24.5	27.4	--	--	24.5
4+ years	26.7	26.3	--	--	26.7
Expected community supervision duration					
1 year or less	--	--	13.0	10.7	13.0
1-2 years	--	--	38.1	34.3	38.1
2-3 years	--	--	18.6	18.9	18.6
3-4 years	--	--	10.8	18.9	10.8
4+ years	--	--	19.5	17.2	19.5
Custody Level					
MI1	8.7	8.8	--	--	4.3
MI2	27.5	24.5	--	--	13.7
MI3	36.1	36.7	--	--	18.0
MED	18.3	21.1	--	--	10.1
MAX	2.0	0.7	--	--	1.1
CLO	7.4	8.2	--	--	4.5
Field Location					
Section 1	--	--	11.8	8.8	11.8
Section 2	--	--	14.7	19.1	14.7
Section 3	--	--	22.1	19.6	22.1
Section 4	--	--	17.8	16.5	17.8
Section 5	--	--	17.8	20.1	17.8
Section 6	--	--	15.9	16.0	15.9
Alternative sentence	7.2	7.1	20.8	30.4‡	13.5
Current sex offender	21.7	19.2	19.2	32.5‡	20.5
Static 99	3.0(0.1)	3.3(2.0)	2.8(0.1)	3.7(1.0)	2.90 (0.1)
High	17.7	18.2	11.2	13.6	15.9
Moderate High	24.0	18.2	26.3	27.3	24.7
Moderate Low	27.8	18.2	33.9	27.3	29.5
Low	30.4	45.5	28.6	31.8	29.9

Stable 2007	10.7(0.2)	12.0(3.1)	8.2(0.3)	10.1(1.8)	9.61 (0.2)
High	41.4	25.0	23.2	37.5	33.6
Moderate	55.4	75.0	57.2	50.0	56.1
Low	3.3	0.0	19.6	12.5	10.2

† Includes 50 "alternate" cases to be used if needed
‡Populations oversampled by 25 subjects

Assessment Training

On February 5th Dr. Hamilton provided a brief training to the team of assessors conducting the pilot. Item level definitions were presented and work sheets for data entry (Excel spreadsheet) were reviewed. In addition, a brief STRONG-R introduction and general interviewing techniques were provided. Assessors were asked to review the definitions and practice the techniques. A second training has been planned for a future date and will include guided examples and role play to further assessors' interviewing techniques. All developed and necessary training materials were provided prior to each training session.

METHODS

Pilot Assessments Collection

Beginning on February 8th WADOC assessors conducted 207 pilot assessments. A total of 143 listed offenders were not assessable during the data collection period. In addition, data collected for seven offenders was incomplete, preventing their use in the analysis. This provided 200 offender assessments for the initial pilot. Of the initial assessments completed 187 (80%) were completed with male offenders and 47 (20%) with female offenders.

A total of four weeks were needed by WADOC assessors to the pilot assessments and complete data entry[2]. An algorithm was created by the WSICJ to transpose (or download) worksheet data into a database to be utilized for analyses. The pilot sample data was then cleaned and recoded to mirror development sample scoring mechanics. This dataset comprised the STRONG-R pilot information that

[2] It should be noted that an additional sample the extra female sample cases required an additional two weeks for collection, processing, and analysis.

would then be used for comparison to development sample projections and current SRA2 risk classification estimates.

Assessors were also asked to provide feedback on the tool. Specifically, a survey was constructed and asked team members to identify items and/or scores that, if modified, would improve functionality of the STRONG-R, assessment timing, and potential barriers for implementation. Survey findings were intended to inform larger training and quality assurance initiatives.

Dataset Descriptions

One intent of the pilot study was to identify potential inconsistencies in coding and scoring of offenders on STRONG-R items/responses. The initial development sample used to create the STRONG-R consisted of roughly 45,000 offenders, assessed prior to their reentry to the community. Offenders reentered the community either from prison or were to receive community supervision directly (without incarceration). All offenders in the initial development sample were released between the dates of August of 2008 and December of 2010. All offenders in this sample had previously been released and were observed for two years to identify instance of new felony convictions.

Since the development of the STRONG-R, additional offenders have been assessed both in the community and in prison facilities. Because the tool is anticipated to be used with inmates, guiding case management decisions, additional data on incarcerated offenders was needed to provide for a more representative sample. Furthermore, additional offenders have reentered the community and recidivated since the initial development sample. To assist the current project, WSICJ amended the development sample with additional offenders, assessments, and recidivism data. This updated development sample consist of roughly 100,000 offenders, of which approximately 70,000 possess recidivism follow-up information.

Finally, the pilot data consisted of 200 offenders randomly drawn and assessed during the study time period. These offenders were drawn from a snapshot of the current WADOC offenders supervised in January of 2016. Given the recency of these offenders within the WADOC system, pilot data does not possess recidivism information.

Analysis Plan

Three primary analysis goals were completed as a part of the study. First, descriptive comparisons between the STRONG-R development sample and the pilot data were completed. This consisted of a detailed item/response analysis, exploring potential inconsistencies between development sample data and pilot data. The intent here was to use pilot data as a proxy for future samples of WADOC offenders and to identify if future samples would score differently than development sample offenders as a result of training or other potential implementation variations.

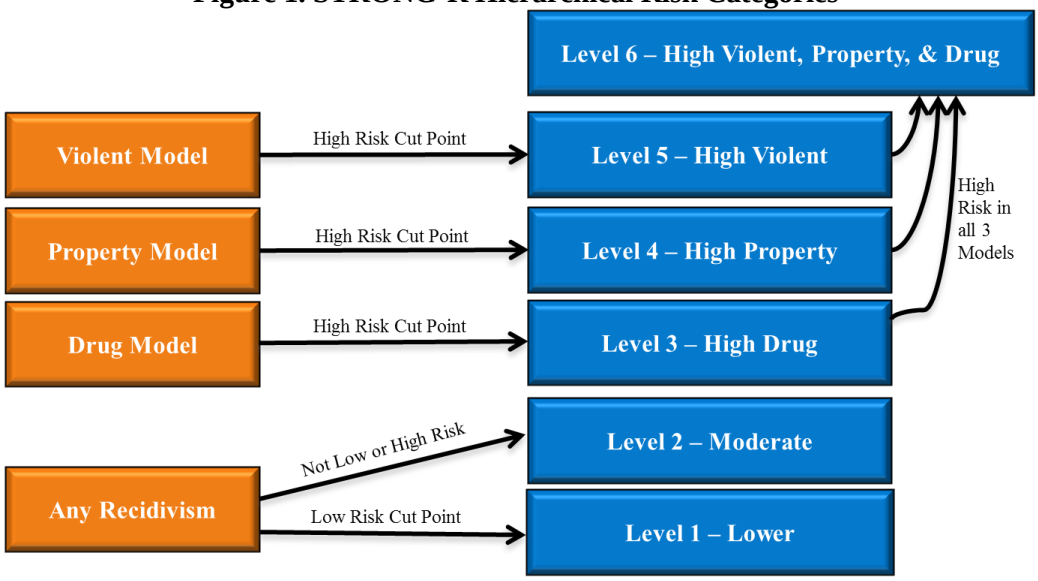
The second study aim sought to create and compare potential cut point options to be used to establish new risk level classification categories. Using pre-established STRONG-R continuous risk scores predicting violent, property, drug and felony recidivism, three cut point options were created. Each cut point option was established in an attempt to balance intrinsic meaning surrounding recidivism base rates while jointly assessing the potential impact on WADOC assessment labor. With regard to recidivism base rates, the updated development sample was used to establish the prevalence of violent, property, drug and 'any' felony recidivism following an offender's reentry. All eligible offenders possess a risk score provided by each of the four recidivism models. Every offender risk score is associated with a probability of recidivism. To construct cut points, recidivism base rates were examined relative to the proportion of offender possessing a probability of recidivism at an established rate. This established rate is set to be greater than the given recidivism base rate for high risk classifications (violent, property and drug) and less than the base rate of felony recidivism for low risk offenders. To provide an assessment of the STRONG-R's impact on labor, cut point options were compared with regard to risk category proportions and recidivism prevalence of the SRA2 score for both development and pilot samples.

Three cut point options were created. The first was created using gender specific male and female samples, setting high risk cut points using recidivism probabilities that were twice the base rate and the lower/moderate risk cut point was set at one-fifth the felony conviction base rate. The second option set high risk cut points at 2.5 times the base rate, while the lower/moderate risk cut point was set at one-half the base rate. The third option was established similar to the first but instead of using gender specific

samples, a gender neutral sample was used to establish recidivism base rates. Again, high risk cut points were set at twice the base rate and the lower/moderate risk cut point was set at one-fifth the base rate for felony convictions. While Options 1 and 2 were set after exploring best fits with the two cut point goals discussed, it was discovered that when gender specific samples were used, female offenders were over classified in high risk categories. Option 3 uses gender neutral samples to reduce this bias and provide a better reflection of offenders' overall risk to public safety.

Categories were assembled based on a WADOC vetted hierarchy of offense seriousness. First offenders with scores that exceed the violent model cut point are identified as Category 5 – High Violent. If the offender exceeds the property model cut point, and has not been indented as High Violent, they are classified as Category 4 – High Property. If the offender exceeds the drug model cut point, and has not been indented as High Violent or High Property, they are classified as Category 3 – High Drug. If the offender has not yet been assigned a category, they are categorized as either Category 2 – Moderate – or Category 1 – Lower – risk based on the felony model. Finally, Category 6 – HVPD – is identified for offers that exceed all three high risk cut points (e.g., violent, property and drug risk scores). Figure 1 provides an illustration of this categorization process.

Figure 1. STRONG-R Hierarchical Risk Categories



Survey

The pilot survey was provided to the 45 assessors to gather stake holder insight and identify potential issues related to the future implementation and training efforts. The survey was created by the Advance Corrections team with additional input provided by WSICJ. All surveys were gathered by the Advance Corrections team and electronic copies were provided to WSICJ research staff. Descriptive data are provided for each item in the results section to follow and the full survey is included in Appendix I.

Race/Ethnicity

Based on concerns of disproportionate minority impact, a breakdown of risk and race/ethnic categories was requested. Senior leadership identified five primary categories of interest. Cross tabulations were completed for the recommended cut point option and the SRA2 by race-ethnicity category.

RESULTS

Descriptive Comparison

Comparing the STRONG-R development sample to the study pilot sample we examined bivariate contrasts of item and response frequencies. While it was not feasible to provide significance testing due to small pilot sample sizes, many item level examinations demonstrated comparatively similar frequencies across the two samples. It should be noted that the pilot sample subjects were identified to have lengthier criminal histories as well as a greater frequency of correctional events. These differences were anticipated due to the greater proportion of incarcerated offenders in the pilot sample.

This comparison also identified items/responses of note due to what is likely inconsistent coding. Some of these inconsistencies may be due to programming or interventions that have increased in use or have not been coded similarly for offenders assessed in prison as compared to those assessed in the community, i.e., offender change programming, vocational programming, Security Threat Group, and annual prison visitations. Other inconsistencies may be related to items that have a relatively lower prevalence in the development sample, i.e., ‘relies on public assistance’, ‘well managed conflict with

partner', and 'IV drug use'. Other inconsistencies may need further examination with WADOC subject matter experts.

Cut Point Assessment

Using the pre-established STRONG-R scoring male and female risk scores were calculated for violent, property, drug and felony models. Model distributions were examined and all were found to be normal, or bell-shaped, with the majority of offenders surrounding the average risk scores, and few with offenders identified in the extreme high and low ends of the distribution. Extreme scores (outliers) were also found to be infrequent and within the normal bounds expected due to random chance.

Base rates were then established for each recidivism type and for both gender specific and gender neutral development samples. Base rates are provided in Table 1. As indicated, 'any' felony base rates are largest, followed by violent felonies for men, drug felonies for women, and an equal base rate is identified for violent, property, and drug felonies in the gender neutral sample. It should also be noted that the violent felony base rate for females is less than half (4%) of that recorded from males (10%) in the development sample.

Table 1. Recidivism base rates for gender specific and gender neutral samples.

Recidivism	Male %	Female %	Gender Neutral %
Felony	24	20	23
Violent	10	4	9
Property	9	8	9
Drug	9	10	9

Male Cut Points

Each of the three cut point options were applied to their respective gender specific or gender neutral development samples. The proportion of each risk category is provided in Table 2. In addition, the SRA2 category proportions are provided as a reference. What is notable is that the category proportions for both Options 1 and 3 resemble that of SRA2. Percentage of cases identified as Lower and Moderate are near identical and when combining the STRONG-R percentages for drug and property as well as the combined violent and HVPD categories one observes similar proportions to the SRA2's high non-violent

(HNV) and high violent (HV) categories. Option 2 does not provide these same similarities, identifying disproportionate category comparisons with the SRA2.

Table 2. Male Cut Points – Risk Category Proportions

SRA2	%	Option 1	%	Option 2	%	Option 3	%
Low	14	Lower	14	Lower	42	Lower	13
Mod	24	Moderate	25	Moderate	22	Moderate	24
HNV	25	High Drug	7	High Drug	6	High Drug	7
		High Property	15	High Property	14	High Property	13
HV	37	High Violent	17	High Violent	11	High Violent	20
		HVPD	22	HVPD	5	HVPD	23

The cut point scores created in the development samples were then applied to the pilot sample. Category proportions were examined and findings are presented in Table 3. Again, Options 1 and 3 illustrate similar proportions, with Option 3 providing estimates that are more in line with the Lower, Moderate and combined high category proportions of the SRA2.

Table 3. Male Pilot Sample Risk Category Proportions

SRA2	%	Option 1	%	Option 2	%	Option 3	%
Low	19	Lower	23	Lower	53	Lower	20
Mod	16	Moderate	20	Moderate	15	Moderate	20
HNV	12	High Drug	4	High Drug	3	High Drug	4
		High Property	7	High Property	8	High Property	7
HV	53	High Violent	27	High Violent	17	High Violent	30
		HVPD	19	HVPD	5	HVPD	19

Female Cut Points

Each of the three female cut point options were then applied to their respective gender specific or gender neutral development samples. The proportion of each risk category is provided in Table 4. Again, the SRA2 category proportions are provided as a reference. The notable contrast between the female and male options is that only Option 3 provides similar proportions as compared to the SRA2. Percentage of cases identified as lower and moderate are similar and the combined drug and property as well as the combined violent and HVPD categories are in line with the high non-violent (HNV) and high violent (HV) categories of the SRA2 than Options 1 or 2.

Recall that Option 3 provides cut point criteria similar to that of Option 1 but was constructed using gender neutral base rates. This option was created to account for the over-classification of female violent offenders that occurs when using Options 1 or 2. With a combined 9% of females identified as High Violent or HVPD, these findings are more in line with the 4% base rate of WADOC supervised females committing violent recidivism, as compared to the combined 39 and 28% identified as High Violent or HVPD using the gender specific samples of Options 1 and 2, respectively.

Table 4. Female Cut Points – Risk Category Proportions

SRA2	%	Option 1	%	Option 2	%	Option 3	%
Low	30	Lower	16	Lower	43	Lower	23
Mod	27	Moderate	29	Moderate	19	Moderate	28
HNV	37	High Drug	5	High Drug	3	High Drug	19
		High Property	12	High Property	7	High Property	22
HV	6	High Violent	21	High Violent	22	High Violent	4
		HVPD	18	HVPD	6	HVPD	5

The female cut point scores created in the development samples were then applied to the pilot sample. Category proportions were examined and findings are presented in Table 5. Again, Options 1 and 3 illustrate similar proportions, with Option 3 providing estimates that are more in line with the Lower, Moderate and combined high category proportions of the SRA2.

Table 5. Female Pilot Sample Risk Category Proportions

SRA2	%	Option 1	%	Option 2	%	Option 3	%
Low	34	Lower	30	Lower	61	Lower	37
Mod	13	Moderate	21	Moderate	5	Moderate	23
HNV	47	High Drug	5	High Drug	5	High Drug	12
		High Property	12	High Property	9	High Property	21
HV	6	High Violent	21	High Violent	16	High Violent	2
		HVPD	12	HVPD	5	HVPD	5

Recidivism Cut Point Comparison

The primary intent of risk categories is to provide discrimination with regard to recidivism. That is, offenders in Lower and Moderate risk categories should possess lower rates of recidivism, while High risk categories should indicate greater proportions of recidivism. Identification of a larger amount of separation, equates to greater discrimination and accuracy.

When examining male recidivism findings between the SRA2 and cut point Option 3 we find better discrimination across the risk categories of the STRONG-R. Specifically, while category proportions were relatively similar (see Table 3), the percentage of recidivism committed is more accurate. That is, the Lower and Moderate categories of Option 3 demonstrate a lower percentage of recidivism for felony offenses. Furthermore, the High Drug and High Property cut points identify a greater proportion of recidivism than the SRA2's HNV category. While only slightly improved, Option 3 indicates a point increase in discrimination over the high violent category of the SRA2. Finally, the HVPD category of Option 3 demonstrates equal-to-substantial improvement across all four recidivism types, further demonstrating the improved accuracy of the STRONG-R as compared to the current SRA2 estimates.

Table 5. Male Recidivism – SRA2-Option 3 Recidivism Comparison

Recidivism	SRA2 Category	STRONG-R Category	SRA2%	Option 3%
Felony	Lower	Lower	7	5
	Moderate	Moderate	20	15
		HVPD		43
Drug	HNV	High Drug	9	17
		HVPD		17
Property	HNV	High Property	9	17
		HVPD		19
Violent	HV	High Violent	17	18
		HVPD		17

Female recidivism comparisons are displayed in Table 6. For female offenders a similar recidivism discrimination pattern is identified when comparing SRA2 proportions with STRONG-R Option 3. That is, while Lower and Moderate risk groups are of similar proportion to the SRA2 categories (see Table 4), the proportion of offenders in the STRONG-R categories committing recidivism is lower. Similarly, the STRONG-R high risk and HVPD categories identify a greater proportion of recidivism when compared to comparable SRA2 categories. Again, these findings identify greater discrimination qualities for the STRONG-R hierarchy classification system.

Table 6. Female Recidivism – SRA2-Option 3 Recidivism Comparison

Recidivism	SRA2 Category	STRONG-R Category	SRA2%	Option 3%
Felony	Lower	Lower	10	5
	Moderate	Moderate	22	15
		HVPD		43
Drug	HNV	High Drug	14	17
		HVPD		17
Property	HNV	High Property	12	17
		HVPD		19
Violent	HV	High Violent	12	18
		HVPD		17

Race-Ethnicity Breakdown

As indicated, five primary racial categories were identified and a breakdown was requested to examine proportions of category assignment based on race/ethnicity. We first provide a descriptives of the five race-ethnic categories. These frequencies are provided in Table 7. One notable finding is that White offenders make up three-fourths of the female population, while that proportion is only two-thirds for male offenders.

Table 7. Development Sample Race/Ethnicity Descriptives

Race-Ethnicity	Male%	Female%	Overall%
White	68	75	69
Black	14	9	13
Hispanic	10	5	9
Alaskan/Native American	3	5	3
Other	6	6	6

Because the STRONG-R is scored via gender specific models, we present the findings of the SRA2 and STRONG-R breakdown by gender. Male findings are presented in Table 8. What is notable is that the risk category proportions as they pertain to the five racial-ethnic groupings is that Lower/Moderate versus High risk categories are very similar when comparing the SRA2 and SRONG-R Option 3. Specifically, the Lower/Moderate risk groups represented 38 and 39% of each tool's overall categorization, while High risk groups represent 62 and 61%, respectively. Using this division between Lower/Moderate and High risk categories one can compare each tool's proportions of racial/ethnic groups. In particular, White-male offenders represent roughly equal proportions in each tool. Similar

findings are identified for Black-male offenders, with a slightly greater proportion identified as Lower/Moderate in the STRONG-R categorization (30%) than the SRA2 categorization (28%). However, when examining Hispanic offenders, a slightly greater proportion identified as Lower/Moderate in the SRA2 categorization (43%) than the STRONG-R categorization (35%). For Alaskan/Native Americans and those male offenders identifying as 'Other' were found to have roughly equal proportions were identified when comparing Lower/Moderate to High risk categories.

Looking at the model effect size (r), we find a small-to-medium effect size for the STRONG-R risk categories (r=.30), while the SRA2 identified a medium-to-large effect size. Both effect sizes suggest substantial variation among risk categories by race/ethnicity. While racial/ethnic disparities are common for nearly all risk assessment instruments, the positive take away is that the STRONG-R reduces disparity when compared to its static-only counterpart – the SRA2.

Table 8. Male Development Sample Risk Category Proportions by Race/Ethnicity

SRA2	White	Black	Hispanic	Alaskan/Native American	Other	Overall
Low	15	6	13	8	21	14
Mod	25	21	30	21	31	25
HNV	28	22	22	21	23	26
HV	32	50	35	50	25	35
Model r	0.4					
STRONG-R Option 3						
Lower	15	7	9	7	19	13
Moderate	24	23	26	22	31	25
High Drug	7	8	7	6	5	7
High Property	15	7	9	11	11	13
High Violent	16	31	22	23	20	19
HVPD	22	25	28	31	15	23
Model r	0.3					

Table 9 provides the Race/Ethnicity breakdown for female offenders. Overall, the Lower/Moderate risk groups represented 58% of the SRA2 and 52% of the STRONG-R tool's overall categorization, while High risk groups represent 42% and 48%, respectively. Notably the STRONG-R categorizes fewer white offenders as Lower/Moderate (53%), compared to the SRA2 (61%). The reverse

is true for Black female offenders, where the STRONG-R classifies a greater proportion of Lower/Moderate offenders (57%) as compared to the SRA2 (49%). However, for Hispanic and Alaskan/Native Americans a slightly greater proportion of Lower/Moderate risk offenders were categorized by the SRA2 (48% & 47%, respectively) as compared to the STRONG-R (41% & 42%).

With regard to the model effect size (r), we find a small-to-medium effect size for the STRONG-R risk categories (r=.30), while the SRA2 identified a medium-to-large effect size. Again, both effect sizes suggest substantial variation among risk categories by race/ethnicity. However, as with the male models, the STRONG-R is again shown to reduce disparity when compared to the SRA2.

Table 9. Female Development Sample Risk Category Proportions by Race/Ethnicity

SRA2	White	Black	Hispanic	Alaskan/Native American	Other	Overall
Low	32	22	21	20	34	30
Mod	29	27	27	27	28	28
HNV	36	35	43	42	32	36
HV	4	17	9	12	6	6
Model r	0.4					
STRONG-R Option 3						
Lower	24	25	16	12	30	23
Moderate	29	32	25	30	23	29
High Drug	20	14	21	20	16	19
High Property	22	14	25	22	22	21
High Violent	3	9	4	8	3	4
HVPD	4	7	9	8	6	5
Model r	0.3					

To further examine racial-ethnic differences the industry standard predictive validity statistic, Area Under the Curve (AUC), was computed for each racial category. Findings of these tests are provided in Table 10. To identify significant differences, confidence intervals (in parentheses) are provided; where overlapping intervals indicate non-significant differences when comparing AUCs, while non-overlapping intervals indicate that model AUCs are significantly different. Given that each model is created with the total (aggregate) sample, we anticipate the 'overall' AUC statistics for each model will differ when

comparing across racial categories, however, one would hope that the differences are rarely significant. A total of 20 model comparison were made (4 models * 5 race/ethnicity categories) and results indicate that only one (Black-Felony) of the 20 comparisons were identified to differ significantly from the 'overall' model. This proportion represents only 5% of the comparisons and is a great indication that they risk assessment models provide accurate scoring for all offender race/ethnicity groups.

Table 10. Development Sample Model AUCs Statistic by Race/Ethnicity

Male	White (CI)	Black (CI)	Hispanic (CI)	Alaskan/Native (CI)	Other (CI)	Overall (CI)
Violent	.72 (.71-.73)	.71 (.68-.72)	.75 (.73-.77)	.70 (.66-.74)	.75 (.71-.78)	.73 (.72-.74)
Property	.77 (.76-.78)	.72 (.70-.74)	.75 (.73-.78)	.74 (.70-.78)	.75 (.72-.79)	.76 (.74-.77)
Drug	.74 (.74-.75)	.73 (.72-.75)	.75 (.72-.77)	.74 (.70-.78)	.80 (.76-.83)	.75 (.74-.76)
Felony	.72 (.71-.72)	**.67(.64-.69)**	.72 (.71-.74)	.70 (.70-.73)	.74 (.72-.76)	.72 (.71-.72)
Female						
Violent	.76 (.73-.79)	.78 (.72-.85)	.81 (.72-.89)	.72 (.63-.82)	.83 (.78-.89)	.78 (.75-.80)
Property	.74 (.72-.76)	.74 (.69-.80)	.73 (.67-.79)	.72 (.65-.78)	.76 (.70-.83)	.74 (.73-.76)
Drug	.72 (.70-.74)	.77 (.73-.82)	.68 (.62-.75)	.74 (.68-.81)	.76 (.69-.82)	.73 (.71-.74)
Felony	.71 (.70-.72)	.73 (.69-.77)	.70 (.65-.75)	.74 (.69-.79)	.73 (.68-.78)	.72 (.70-.73)

Survey Results

One goal of the pilot assessment was to receive feedback from the assessors concerning the process and the assessment itself. As part of the process, case managers were asked to complete a nine question survey to determine barriers, length of time invested in the assessment process, and other feedback. Most assessors completed a survey for each assessment they completed however, this made their answers repetitive. The east side of the state completed one survey per assessor with a total of 122 surveys completed. The list of survey items is provided in Appendix II. The current section provides summary descriptions of survey findings.

Time Investment

Overwhelmingly, assessors indicated that they invested little file review time in preparing for the interview, and spent at least 45 minutes to one hour conducting the face-to-face interview with the offender during the assessment process. The assessors agreed that with practice and with growth in their

familiarity with the tool, the interview will require less time and estimated that the assessment interview could be completed within 45 minutes with a fully participating offender. Almost every time the assessor examined the time it took to complete the assessment they indicated that the time taken to conduct the offender interview was combined with the time taken to complete the assessment. Meaning, the assessment was scored during the conversation with the offender for clarity.

Confusing or challenging questions

Overall, most assessors indicated that there were no confusing or challenging questions. However, for responses other than no, the feedback was quite diverse when assessors were asked about confusing or challenging questions. The assessors stated that when it came to employment questions, it was difficult to score the answer if the offender was employed while in Work Release. Further, concerns were expressed that when an offender is serving a sentence in prison versus being confined in prison on a violation, it was difficult to score without the ability to note the difference. A few additional concerns raised were including a better definition of clean and sober, repetitive wording in questions, having to rephrase for some offenders, difficult to identify which behavior to address, and some questions required some assessor assumption. Lastly, there were multiple questions that at least one assessor mentioned specifically. Of the 122 surveys, only 20 of the 1016 items were listed as confusing in some way (and few were mentioned by more than one survey), which include: 30 (1), 31 (1), 32 (1), 36 (1), 41 (3), 43 (3), 47 (2), 48 (3), 63 (2), 65 (2), 66 (4), 67 (2), 68 (3), 77 (3), 85 (1), 89 (3), 90 (1), 91 (1), 97 (1), and 103 (1). These items should be further explored to ensure item level definitions and training help reduce confusion going forward.

Level of confidence and additional resources needed

The skill of the staff is of a high standard. Most indications were that their level of confidence in conducting and scoring the interview are Confident, Very confident and a handful of Expert. Overall, most assessors indicated that they would not need any additional resources to complete the tool. A few

assessors would like to assure that they are able to view any applicable Pre-Sentence Investigation report and have a detailed criminal history of the offender. Lastly, a few additional resources that were mentioned were a simple worksheet, the file review, TX reports, a question pertaining to teenage crime or crime not on record, CD assessments, brief definitions, and a criminal history description.

Particular groups that the tool will be more/less effective assessing

Assessor feedback included sex offenders, the mentally ill and low functioning offenders, first time offenders, drug offenders including causes of drug use, offenders identified as a security risk, and long-term prison offender. The reasons these types of offenders were identified were not explained. The Advance Corrections team will follow-up with assessors concerning this information gap and try to clarify their concerns and identify possible solutions.

Additional outputs or information that the assessment could provide

Almost every assessor left this question blank however, one assessor pointed out that according to research, impulsivity is a criminogenic need correlated with crime and violence, and perhaps the variable should be expanding under attitudes/behaviors. While the majority of the assessors did not have input to this question there were a few requests to develop a job aid that defines each questions of the tool. A few additional concerns raised were factoring in past violations, programming for offenders with long sentences, needing a 'Not Applicable' option for questions, and handouts for offenders providing information about the assessment.

Conclusions and Recommendations

While limitations of sample size have been discussed throughout, the pilot sample of STRONG-R assessments has provided preliminary findings that indicate the likely future successful implementation of the tool. If cut point Option 3 is selected, one can expect similar risk category proportions as the SRA2, which should help stem concerns of increased or altered supervision labor as a result of the STRONG-R's implementation. There is also an intrinsic meaning behind the cut point selection, deriving base rates

calculations of the WADOC populations as they relate to probabilities of future recidivism indicated by offenders' risk scores.

Findings also demonstrate that model STRONG-R cut points will provide more accurate estimates of risk over the current SRA2. As indicated, category discrimination is improved when using the STRONG-R, placing fewer recidivists in Lower and Moderate risk categories and a greater proportion of recidivists in higher risk categories. The addition of the HVPD risk category will also provide a more comprehensive understanding of what the literature describes as "opportunistic offenders" (McGloin, 2012).

We also note the distinction of the gender neutral versus the gender specific cut points. While the distinctions in the male samples were slight, if one were to assign risk categories based on a gender specific sample for females, the result would be an over-classification of female offenders as it pertains to their overall risk to public safety. In contrast the gender neutral sample cut points provide more realistic estimates of female risk (especially for violent offending) and category proportions are more similar to that of the SRA2.

Comparing the SRA2 to the STRONG-R risk categories by race/ethnic groupings revealed reductions in disproportionality of risk classifications across race/ethnic categories; however, substantial variation still remains. That said, accuracy of the risk models remained consistent, finding only one instance (in 20) of a significant Area Under the Curve (AUC) variation from the aggregated development sample. Furthermore, survey findings by assessors generally revealed confidence in the tool's use and the duration of time needed to complete the interview and scoring. Several notable items were identified to be addressed for training prior to implementation

Next Steps

There are several next steps needed following the conclusion of this pilot study. First additional assessments of categorical and prediction accuracy differences ofmore specified populations (i.e., sex offenders and sentencing alternatives) will provide knowledge of the STRONG-R's impact on current policies that surround these offender types. Second, several items/responses were identified to have

inconsistent scoring patterns in the pilot data. These items/responses should be further tracked and assessed by subject matter experts to outline how best to train and implement the tool to provide consistent, reliable scoring.

The current consulting contract outlines an assessment of interrater reliability to be examined next. In addition, training and QA modules will be developed based on the lessons learned from the pilot data, survey findings, and additional WSICJ expertise. Finally, needs assessment models are currently being examined to identify potential intervention linkages and service gaps.

APPENDIX II. Assessment Survey

Date: _____

Assessor's Name: _____

DOC# of offender assessed: _____

1. Length of time taken to review the case file?

2. Length of time taken to conduct the offender interview?

3. Length of time taken to complete the assessment?

4. Were there any confusing or challenging questions for the Interviewer?

5. What was the level of confidence, on a scale of 1-5 of the scoring for the interviewer? 1=not confident, 2=somewhat confident, 3=confident, 4=very confident, 5=expert (circle one answer).

6. With practice, would you expect the assessment to be completed quicker? If so, what would you anticipate your average length of time to complete?

7. Are additional resources needed to assist with tool completion? If so, what should be provided?

8. Are there particular groups that you feel the new tool will be more/less effective assessing?

9. Are there additional outputs or information that the assessment could provide (in addition to RLC) that would help your assessment, management, and case planning of offenders?