cdps
COLORADO
DEPARTMENT
OF PUBLIC SAFETY

Division of Criminal Justice
Jeanne M. Smith, Director

700 Kipling Street
Suite 3000
Denver, CO 80215-5865
(303) 239-4442
FAX (303) 239-4491

To: Pete Weir, Director CDPS
From: Kim English, Director ORS
Date: October 4, 2009
Re: External Evaluation of the Colorado Actuarial Risk Assessment Scale, Ver. 5

Following the re-development and re-validation of version 4 of the Colorado Actuarial Risk Assessment Scale (CARAS), the Office of Research and Statistics sought external reviews of the methods and processes undertaken for this effort. This Ver. 5 update, as required by legislative mandate, was completed in November 2008 by actuarial scale consultant Marshall Costantino of Analysis, Research and Design, Inc.

Two experts were recommended to conduct a review of the CARAS. Gerald G. Gaes, Ph.D. is a Research Faculty member at the Center for Criminology and Public Policy Research at Florida State University and former Director of Research (1988-2002) and Researcher (1980-'86) at the Federal Bureau of Prisons (1980-'86). S. Christopher Baird is the Executive Vice President of the National Council on Crime and Delinquency directing the Children's Research Center for that organization since 1985. Please see the attached reviewer bio document for additional background information for the reviewers and Mr. Costantino.

In brief, Dr. Gaes indicates that Mr. Costantino employed techniques that "are quite systematic and technically proficient" and that the ability of the CARAS to differentiate recidivists from nonrecidivists is "quite good." Gaes summarizes by stating, "Taking everything together, this was a competent and well executed scale development process…." The complete review by Dr. Gaes is attached.

Mr. Baird, who reports being involved with risk assessment research since 1972, introduces his review by stating that he found the CARAS-5 "quite innovative." He concludes by offering, "The CARAS-5 is a well constructed scale with discriminatory power rivaling (and for the most part, surpassing) other risk models used across the nation." The complete review by Mr. Baird is attached.

Both reviewers offered valuable advice for future versions and further development of the CARAS. The Office of Research and Statistics and consultant Mr. Costantino are studying these suggestions to further improve the performance of the CARAS.

Reviewer and Consultant Bios
Colorado Actuarial Risk Assessment Scale, version 5

Gerald Gaes, Ph.D.
Gerald Gaes, Ph.D. received his doctorate in social psychology from the State University of New York at Albany in 1980. He worked for the Federal Bureau of Prisons for 20 years, including as the director of the Office of Research. He served as a visiting scientist at the National Institute of Justice for five years; he also served a two-year detail at the United States Sentencing Commission. Gaes is first author of Measuring Prison Performance: Government Privatization and Accountability; he has published extensively in professional journals, including Crime and Delinquency, Criminal Justice Review, Criminology and Public Policy, Justice Quarterly, and Punishment & Society. In July 2000, he received the U. S. Department of Justice Attorney General's Distinguished Service Award for the correctional research that he conducted during his BOP career. (Complete vitae as of August 2009 available upon request.)

S. Christopher Baird
Christopher Baird is the Executive Vice President of the National Council on Crime and Delinquency/Children's Research Center and has directed the Midwest Office in Madison, Wisconsin since 1985. He has designed risk assessment, classification and case management systems for child welfare, adult probation and parole, and juvenile justice systems. He developed and managed the National Institute of Corrections Model Probation and Parole program which was implemented in 31 state agencies and hundreds of county probation departments throughout the United States. Mr. Baird served as principal investigator on two grants from the National Institute of Justice, including a comprehensive evaluation of the Florida Community Control Program. From 1990-1997, he directed NCCD's Children's Research Center which developed risk assessment and decision making systems used in Child Protection Services for over 50 state and county agencies in the United States and Australia. He and colleagues wrote a comprehensive evaluation of the system in Michigan assessing its impact on subsequent abuse and neglect. He directed and authored a national study funded by the Office of Child Abuse and Neglect (OCAN) that compared child protective services risk assessment systems in four jurisdictions. He is currently conducting research for the Casey Foundation's workforce initiative.

Mr. Baird has authored numerous journal articles and other publications on research, program development and management issues in child welfare, juvenile justice, and corrections. In 1992, he received the University of Cincinnati Award from the American Probation and Parole Association for outstanding research contributions to the field. In 2001, he and his colleague Dennis Wagner received the Pro Humanitate Literacy Award for "The Relative Validity of Actuarial and Consensus-Based Risk Assessment Systems" from the North American Resource Center for Child Welfare. In 2004, he received the Grace B. Flandeau Aware for his contributions to child welfare. His educational background includes a Masters degree in Economics.

Marshall Costantino

Marshall Costantino's academic training includes a Bachelor of Science degree in Applied Mathematics and a Master of Science degree in Quantitative Economics with a minor in Operations Research. His business and government experience varies among the consumer and commercial credit industry, Federal Government procurement financial analysis, worker's compensation insurance and criminal justice. For the last 33 years, he has been involved in at least 50 consumer and/or commercial credit scoring developments and implementations. During his 12 years with Citigroup, he headed departments participating in credit policy development and implementation for a number of start-up and turnaround businesses. Citigroup had a number of quantitative financial risk management groups whose charge was to identify potential bad accounts and to administer quantitatively based credit policy. Because of his work with these groups, Marshall was placed on a team of internal consultants who were called upon to "put out fires" anywhere in the world at any time.

Marshall formed Analysis, Research & Design, Inc. in 1981 in order to take advantage of the consulting opportunities that arose as a result of guest speaking engagements at conferences all over the world. Between 1981 and 1987, AR&D was a part-time venture but in 1988 it became a full-time business. Since then, Marshall has split his professional time between consulting and teaching undergraduate and graduate courses at various colleges and universities in the Denver area. Currently, he serves as adjunct faculty at the University of Denver's University College. He has taught courses ranging from Business 101 to graduate level economics, statistics, finance, operations research and legal compliance.

Eleven years ago, Marshall built the first claim scoring system for Pinnacol Assurance, Colorado's quasi-government/private partnership in worker's compensation insurance, saving Pinnacol Assurance $56,000 in prevented fraud losses during the first six months. Over the past 11 years, he has analyzed the legal function, sought to identify medical provider fraudulent transactions using Benford's Law, and devised four policy renewal scoring systems. Over the last 10 years he has been retained by two sub-prime auto lenders.

Visit ARD's website to review some of the major projects in which Marshall's agency has been involved, at http://www.arddenver.com .

Review of CARAS, Colorado Actuarial Risk Assessment Scale

Gerald G. Gaes

Florida State University, Criminal Justice Consultant

September 9, 2009

Review of CARAS, Colorado Actuarial Risk Assessment Scale

Marshal Costantino, Analysis, Research and Design, Inc., was contracted to develop the new CARAS instrument. In this report, I review the scale development methods and the statistics that indicate how well the scale performs. Most of my comments in this review are based on the document entitled "CARAS Information Request", a follow up memo from Costantino entitled "Clarification Response to your Questions on the DOC Information Request" in which he replied to questions I had about the original document,. I also used a spreadsheet Costantino sent showing how the weights for CARAS were computed, and other documents which were primarily memos indicating progress and decisions made throughout the CARAS development.

Recidivism Definition and Technical Violators

One of the crucial decisions in the scale development was the definition of outcomes for the purpose of classification. Under guidance from the Division of Criminal Justice, Colorado Department of Public Safety (DCJ), Costantino divided the post-release events into returns to prison based on a new felony filing (call these recidivists), returns based on technical violations, and people who remained in the community (call these non-recidivists). The rationale behind this decision is that DCJ considers recidivism to mean harm to members of the community and revocations due to technical violations are due to failures to follow terms of supervision. I suspect that the public policy goal of returning some technical violators to prison is to preclude crimes that they may commit. Therefore, as noted by Costantino, they are removed from the pool of parolees possibly prior to their committing a new crime. Since we cannot know whether they would have committed crimes had they continued to technically violate conditions of their supervision (or perhaps were committing crimes but had not been arrested), this is a difficult problem to handle. Most analysts treat returns to prison based on a technical violation as the same event as a return for a new felony filing. There are different ways to examine the relationship between the classification predictors and the different outcomes. One could use a form of regression, multinomial regression to see if the same factors that predict returns for new felony filings are similar to returns for technical violations. Another way to model the prediction of failure is to use survival methods and treat technical violations as a censoring event for returns based on felony filings. Then one treats technical violation returns as the event of interest with returns to prison based on a felony filing as a censoring event.

While these use regression to estimate the predictive power of variables, Costantino chose the more traditional way of classification, and what he did is consistent with the way most analysts develop classification tools in criminal justice. Costantino chose to model the parolees returned to prison based on 1 or more felony filings relative to those who remained on parole with no felony filing

excluding people who were returned to prison for a technical violation. The document entitled "CARAS Information Request" provides the following explanation:

> Analysis performed during the development process indicated that 2/3rds of the technical violators were more similar to and could be combined with the nonrecidivists and 1/3rd of the technical violators were more similar to and could be combined with the recidivists for the purpose of developing the final scoring table. Were the scoring table developed using only the information on the known recidivists and nonrecidivists, it would misrepresent the actual population to which it is to be applied. Similarly, had all technical violators arbitrarily been combined with the recidivists, the scoring scheme would misrepresent those who appear more similar to the nonrecidivists. (p. 2)

Once Costantino had completed the development of the classification scale excluding the technical violators (TV's), he compared the TV's to the recidivists (R's) and the non-recidivists (NR's). This comparison is documented in Appendix A of the "CARAS Information Request" document. On page 3 of the report it indicates that the distribution of TV's into the 5 categories of risk (very low, low, medium, high, very high) was very even. It was argued that if TV's looked more like recidivists, then their risk category distributions would look more like recidivists than non-recidivists. In fact, it looks like neither. This indicates to me that the characteristics of technical violators as summarized by their risk scalar value seems to be midway between the recidivists and non-recidivists, where the classification development was based on returns to prison based for felony filings.

I have dwelled on this part of the scale development because it is a major concern of the DCJ. It is clear to me that their choice to use only the returns to prison based on felony filings was a valid choice.

Development Method

Variables. The techniques Costantino used to develop and validate the risk classification system are quite systematic and technically proficient. The sample sizes are large. He used a development and validation sample. The pool of automated risk predictors is also quite large and very comprehensive. The fact that the Level of Supervision Inventory (LSI) is only one of a host of potential predictors culled from their automated system shows the richness of the pool of predictors. The LSI does include criminal history information in its scoring, yet Costantino demonstrates the prediction can be improved with other risk factors including criminal history elements. Also, the fact that all of these are automated elements means that there is a built in efficiency to the process. No extra paperwork has to be done.

I looked closely at the variable list in Appendix C. The factors are quite comprehensive including socio-demographic, extensive criminal history including juvenile record, psycho-social factors, mental health, education, peer relationships, drug history, attitudes and emotions, employment prior to prison, characteristics of the current offense, and conduct in prison.

The technical steps in the scale development are outlined in Appendix D. I asked Mr. Costantino for clarification of some of these steps and he provided me with a lengthy exposition which was extremely helpful. In fact, his memo could be incorporated into any further CARAS development documentation.

Post-release period. After getting the data, the first major step for Costantino was to determine an optimal time frame for the recidivism analysis. Costantino used procedures to choose the most optimal post-release period to develop the risk classification procedure. He selected a prediction set of variables that could be used for a 2,3,4, or 5 year post-release time frame. He then evaluated the percentage of the correct classification of parole successes excluding the TV's. He also evaluated the overall correct classification of the prediction set over time and found that the highest percent of correct classification for both successes and failures occurred at three years.

Missing data. Missing data were handled in an appropriate way similar to the method the U. S. Census Bureau uses in its "hotdeck" procedure. Because missing data could be present when CARAS is used to score offenders, DCJ needed a system to incorporate missing data into the final score. An analyst could have used modern missing data imputation procedures to do the scale development; however, when missing data did occur as staff tried to use the new scale, this would have been a big problem and would delay classification until the data were entered.

Predictor selection. To select items from the pool of 177 as the best predictor set, Costantino describes using divergence tests composed of standardized difference test in the mean values for the recidivists and non-recidivists. This allowed him to pare down the original large pool of predictors into 25. He then used a discriminate analysis procedure entering all 25 variables simultaneously to uncover predictors that were highly correlated. He then did a stepwise discriminate analysis to see how individual predictors affected the both the discriminant power of other variables and the overall correct classification rate. This led to the final set of 9 variables.

Costantino translated the 9 predictors into Rate Increase Factor form. He applied this weighting to the recidivists, non-recidivists and technical violators. Even though the scale development excluded the TV's, he wanted to see how they would score on the CARAS to compare their "risk" levels to the

recidivists and non-recidivists. This is where they appear to be somewhat more like the non-recidivists although they are equally distributed over the five risk categories – very low to very high.

Costantino compared logit and OLS regression to see which produced better weights for purposes of classification. The logit estimation procedure was better. Using the logit weights, he reclassified the TV's into recidivists and non-recidivists based on their classification result. As I have already said, although this was not the only way to approach this problem, however, this was a reasonable way to address the issue.

The final scale values were based on translating the coefficients for the logistic prediction equation. A "Rate Increase Factor" (RIF) was based on forming the ration of the recidivism percentage in a given category to the lowest ranked category for a given variable. For example, the variable arrested under the age of 16 has a no and yes category. The lowest recidivism percentage is the "no" category and this receives a RIF of 1.0. Comparing the recidivism percentage of those who had an arrest prior to 16 (47.1%) versus those that did not (38.75%) produces the ratio 47.1/38.75 = 1.46. This is the Rate Increase Factor. The logit coefficient for this variable was 1.13. The final weight was the logit coefficient times the Rate Increase Factor times 10 rounded to the nearest integer. In this case, that produced a weight of 11 for those offenders who did not have an arrest under the age of 16 and a weight of 17 for those that did. The intercept was also weighted as well. Using this composite set of weights, an offender can get a score between 1 and 79.

Classification. The acid test of a classification system is how well it discriminates between recidivists and non-recidivists and the extent to which people are not being correctly classified. CARAS has a correct classification percentage of about 71 percent and receiver operating characteristic AUC (area under the curve) of .76. This latter measure is a summary of correct classification to incorrect classification. These are quite good. The scale divergence criterion is also good, close to a 1 unit standardized difference. Of course, future validation will insure that population characteristics may not change the scale validity values.

The "CARAS Review Request" also has a short discussion on reliability. This is the psychometric notion that different people will score the scale the same way (inter rater reliability) or that if the scale is measured over time on the same person and the risk items do not change, one will get the same scale result. I think the best way to handle inter rater reliability is to do auditing of the data entry to insure it is being done correctly.

There was a special discussion of CARAS's ability to classify violent and sex offenders. On pages 14-17 of the document "CARAS Information Request," there are data on how each of the scale items compares for the violent versus non-violent subgroups and the sex offender versus non sex

offender subgroups. CARAS rank orders these subgroups correctly for each of the scale categories very low to very high risk. There is very little difference between the average scores on each of the scale items for the violent versus non violent, and sex offender versus non sex offender subgroups. The correct classification levels are comparable for the sex offender and the overall sample on which CARAS was developed and validate. The correct classification percentage for the violent subgroup may be slightly better than the overall sample. These data show the validity of CARAS for predicting whether someone will return to prison for a new felony filing whether they are violent offenders or sex offenders.

Summary

Taking everything together, this was a competent and well executed scale development process and the fact that this is based on automated items makes it an efficient process going into the future.

Future steps. One issue DCJ should be aware of for future scale development is that criminologists are now questioning how risk classification and the criminal justice response affects outcomes. There is a seminal paper by Bushway and Smith, (2007). The argument is that risk classification and other predictive criminal justice tools are not used in a vacuum. So that people who get a high risk classification may have their parole delayed or may have closer supervision when released to the community. The former may decrease recidivism (age, maturation effect). The latter may increase the possibility of a technical violation (closer scrutiny, more conditions of supervision). This is a complication most analysts are ignoring when they do scale development and it is a very difficult issue. I thought that it is important to point out that criminologists are beginning to tackle this problem and that DCJ should be aware of the issue for future scale development. There are also a host of new techniques that are being experimented within criminal justice to classify populations. One important technique is a recursive partitioning procedure called classification and regression trees (CART). CART has been used by Berk (2008) to uncover classification rules for quite rare criminal justice events. Again, this and similar tools are cutting edge and are not yet widely accepted. Nor are there readily available tools to do these analyses. CART is available in SPSS, but not some of the additional tools to refine the classification tree. There are procedures implemented in the R statistical set of packages. The advantage to regression trees is that it can reveal complicated underlying relationships between variables that are not readily revealed with standard classification procedures.

References

Berk, R. A. (2008) *Statistical Learning from a Regression Perspective*, New York" Springer

Bushway, S., & Smith, J. (2007). Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us. *Journal of Quantitative Criminology* , 377-387.

M E M O R A N D U M

to: Kim English, Ph.D., Research Director, Colorado Division of Criminal Justice
from: Christopher Baird, Executive Vice President, National Council on Crime and Delinquency
subject: Review of CARAS-5
date: September 3, 2009

Thank you for the opportunity to review all of the analyses conducted to develop the CARAS-5. Prior to discussing the CARAS-5, I want to provide a short summary of my experience.

I have been involved in risk assessment research since 1972, working in adult corrections, juvenile justice, and child welfare. Over the last 20 years, Dr. Dennis Wagner and I have completed over 50 development and validation studies of risk instruments. The risk assessment model we developed for child welfare, Structured Decision Making®, is the most widely used case management system in the world. In 1980, I developed a probation and parole risk assessment system for the National Institute of Corrections. As recently as 2001, a National Institute of Justice survey found that this system was still used by 60% of the probation and parole agencies that responded to the survey. Our research on risk assessment has garnered several national awards over the years.

Overall, I found the research supporting the CARAS-5 to be very solid; in some respects it was quite innovative. Established research protocols were used and all of the statistical methods employed were appropriate. The study cohort was large (5,850 cases) and was appropriately divided into development and validation samples. The follow-up period (36 months) is actually longer than what is found in most studies and adds to the strength of the analysis. The level of discrimination attained between risk groups was excellent, rivaling anything the National Council on Crime and Delinquency (NCCD) has developed or reviewed. I should note that it is always preferable to use a risk instrument developed for a specific state's population, rather than importing a generic system such as LSI or COMPAS. The CARAS-5 will undoubtedly outperform such models.

There are two issues that are critical to evaluating the efficacy of a risk assessment instrument. First, it is important that each risk level contains enough cases to make each designation meaningful. The dispersion of cases across CARAS-5 risk levels is quite good, ranging from a low of nearly 13% for the very low risk category to 31.6% for the very high risk category. Second, recidivism rates observed should increase significantly as risk levels increase. The "spread" attained for CARAS-5 (17.2% to 76.1%) is very impressive. It is also impressive that the spread attained was replicated in the validation sample.

Separate analyses were conducted to test the utility of the CARAS-5 in assessing risk for violent offenders and for sex offenders. The CARAS-5 does very well with both groups. As expected, violent offenders have a lower overall rate of recidivism than property offenders—about half of all violent offenders rate low or very low risk, with a combined recidivism rate of under 15%—while the recidivism rate for very high risk violent offenders was nearly 74%. The level of discrimination attained for sex offenders was somewhat lower, but still impressive, ranging from 15% for the lowest risk group to 62% for the highest risk group.

Technical violators (TVs) were treated differently than what NCCD usually encounters. However, I feel the actions taken were not only appropriate, but innovative. TVs were omitted from the initial scale construction effort, based on the fact that they were neither "failures"—no new crime was reported—nor were they successes, as they had been returned to prison following parole. Omitting these offenders from the initial analyses allowed for the development of a pilot risk instrument based on cases that were either "true failures" or "true successes." Scoring the TVs on the pilot scale revealed that most (two thirds) fell into the lower risk levels and therefore were likely "successes." This finding may indicate that Colorado parole officers are initiating revocations too quickly. This would correspond with trends NCCD has seen in the other states where rates of technical violation have increased in recent years. To the extent that this adds to the time that offenders representing little risk to public safety spend in prison, it is a misallocation of resources.

TVs were added to the analysis to derive the final scale. While the manner in which TVs were added is different than the typical approach used by correctional researchers, I am convinced that the integrity of the analysis was preserved and that there was minimal impact on the model's ability to correctly classify offenders into different risk levels. In fact, given the high rate of technical violations, I believe that the approach taken in the Colorado analysis is superior to simply categorizing all TVs as recidivists (an approach frequently used by other researchers).

NCCD does have three recommendations. First, while the 2002 validation sample indicates that the CARAS-5 is quite robust (that is, it will work well across populations and perhaps over time) further analysis would prove beneficial. Given the potential value of the instrument in assisting the parole board and parole officers with public safety issues, it would be wise to further validate the CARAS-5 using release cohorts from 2003, 2004, and 2005. These cohorts would provide a minimum of a three-year follow-up after release from prison and test the instrument's validity on more recent parolees. NCCD sees this step as critically important. If the CARAS-5 works well with these populations, its validity cannot be questioned, and the results should engender greater confidence in those who use the instrument to assist with decision making.

Second, we suggest further study of technical violators. It is clear that a significant number of low risk parolees are being returned to prison for technical violations of parole. Steps could be implemented that would enhance parole success rates and lower the cost of corrections. Such steps could include the following:

- Identifying lower risk parolees who are most at risk of a technical violation and alerting parole officers so that proactive actions can be taken;

- Providing parole with a system of graduated sanctions that keeps these offenders in the community whenever possible;

- Training officers to more effectively supervise these offenders in the community; and

- Training parole officers to use sanctions other than prison when possible.

Finally, while the CARAS-5 is valid across racial groups and for female offenders, there are some "overlaps" in the recidivism rates by risk level between groups. For example, low risk males had a recidivism rate of 23.5%; moderate risk females had a recidivism rate of 22.6%. Thus, these two groups, in terms of their risk of recidivating, are very similar, but the risk labels attached to each group may result in different actions by either the parole board or parole officers. These issues are easily addressed, either by policy or changes in risk labeling that better reflect the base expectancy rates established for each subgroup. One possible solution for female offenders is presented in the table below.

Risk Group	Recidivism Rates		
	Males	Females	Possible Solution: Female Risk Categories
Very Low	18.0%	7.3%	Very Low
Low	23.5%	18.5%	
Medium	33.6%	22.6%	Moderate
High	46.8%	36.5%	
Very High	76.1%	76.3%	Very High

Combining risk groups for females into three categories—very high, moderate, and very low—would provide greater equity and eliminate "crossover" of recidivism rates among risk groups. Obviously, there are other possible solutions to these issues, but to ensure equity, the "overlap issue" should be addressed.

Summary

The CARAS-5 is a well-constructed scale with discriminatory power rivaling (and for the most part, surpassing) other risk models used across the nation. Great attention was paid to details often overlooked by researchers, and the developer introduced a creative method for dealing with technical violators. The CARAS-5 has substantial value to decision makers in Colorado and should be quickly validated using more recent release cohorts. NCCD's experience suggests their instrument will prove robust over time in Colorado.

Please feel free to contact me with questions or if additional information is needed. It was a pleasure to work with your development team and to review work of this quality.