Response to Thomas Morton's "Risk Wars" and "An Alternative View of Structured Decision Making Research"

When research is critiqued, it is usually done in a thoughtful manner. Deliberate misrepresentations are rare and present a real conundrum. To respond can bring undeserved attention to the critique, while simply ignoring the attack can result in inaccurate perceptions that remain in place for years. Thomas Morton's recent publications, "Risk Wars" and "An Alternate View of Structured Decision Making Research," present such a conundrum.

We have chosen to respond because these publications represent a significant breach of professional ethics. Mr. Morton has deliberately misquoted and misrepresented the work of the Children's Research Center (CRC). His proclivity to misquote and misrepresent the work of others is well established. In "Risk Wars" he misquotes a paper written by Judith Rycus and Ronald Hughes. In another essay, "Is Cause a Bad Word," he attributes the words of others to Dennis Saleeby. Responses to both essays have clearly identified these transgressions. Mr. Morton also has a long history of misrepresenting work of CRC, including public allegations that CRC risk instruments are "racially biased," an inaccurate charge that has been refuted in three studies. To illustrate the level of misrepresentation in Mr. Morton's essays, a few examples from "Risk Wars" and "An Alternative View of Structured Decision Making" are presented below. A point-by-point response to the latter essay follows these examples.

In "Risk Wars" Mr. Morton states, "Actuarial models never classified maltreatment according to the seriousness of associated harm." This constitutes a major theme of his argument and is presented as a serious shortcoming of actuarial risk assessment.

Shortly after Mr. Morton distributed the "Risk Wars" paper at a National Resource Center on Child Maltreatment (NRCCM) sponsored meeting in Wyoming, a former Deputy Director of Social Services for the State of California wrote to Mr. Morton informing him this statement was in error. The California research included outcomes of injury, serious injury, and

subsequent placement, tracking each of these outcomes over a 24-month follow-up period. In total, seven CRC studies include placement and injury as outcomes; and in general, CRC risk instruments discriminate better on variables that reflect serious maltreatment than they do on recurrence alone.

Mr. Morton admits his error in a letter sent to county child welfare directors in Ohio stating, "We accept, as one might expect, the incidence of injury and placement is higher in higher risk classifications." This directly contradicts several statements in "Risk Wars," yet Mr. Morton does not feel ethically bound to issue a retraction or correction. Instead, he continues to distribute the paper and it remains uncorrected on the Child Welfare Institute (CWI) website.

In his letter to Ohio directors, Mr. Morton attempts to explain his misstatement with the following: "Our point is that risk classification, by itself, is not sufficient for safety decision making." Since no such claim has ever been made by CRC, his point is irrelevant. What is relevant is that Mr. Morton feels it permissible to support his "point" with statements he knows are false.

In his November commentary, "An Alternative View of Structured Decision Making," Morton not only misquotes CRC's work, he completely fabricates a statement and attributes it to CRC researchers. First, the misquote: Mr. Morton writes, "Baird and Wagner (2000) argue that classification is 'a clear indication that such families may require more attention and more services because cases in this designation tend to fail at higher rates than cases in other classifications.'" This sentence makes no sense and indicates that something is seriously amiss. The correct quote is: "Classification recognizes that a <u>high risk designation</u> (emphasis added) is not a prediction of failure. It is, instead, a clear indication that such families may require more

attention and more services because cases in this designation tend to 'fail' at higher rates than cases in other classifications (pp. 851-852)."

Omitting our reference to high risk cases allows Mr. Morton to combine the moderate and high risk categories to make his point. To complete his argument, however, requires further distortion. He states, "The authors suggest that all families that score above one standard deviation below the mean should receive a government intervention (Baird and Wagner, 2000, p. 885)."

Our article ends on page 871 (volume 22, numbers 11 and 12, *Children and Youth Services Review*). There is no page 885. The bigger problem is that nowhere in the entire article is there anything that even remotely resembles such a suggestion. Mr. Morton's statement is pure fabrication. Without these two serious distortions, his point that 84 percent of substantiations would be opened for services is totally without foundation.

While it is easy to dismiss most of Mr. Morton's "points" as distortions intended to only promote his own interests, deliberate misrepresentations such as those noted above should not be dismissed by the child protection field. They represent clear violations of any standard of ethical behavior.

CRC's Point-by-Point Response to "An Alternative View of Structured Decision Making"

Morton: *"At a recent meeting a colleague implored others in the group to 'make them show you the research.' The person, a supporter of Structured Decision-Making (SDM) developed by the Children's Research Center (CRC), seemed to be saying that the research supporting SDM is definitive and conclusive. While the findings may be encouraging for its developers, they are hardly definitive."*

Response: No one has ever claimed that SDM™ research is definitive. Research in the social sciences rarely is. Mr. Morton was simply being asked to produce some data to support the approach he recommends.

Morton: *"CRC offers three reports on its website supporting the validity, reliability and efficacy of SDM. These reports are titled: 'Child Abuse and Neglect: Improving Consistency in Decision-Making.' 'The Michigan Department of Social Services Structured Decision-Making System – An Evaluation of Its Impact on Child Protective Services,' and 'Evaluation of Michigan's Foster Care Structured Decision-Making Case Management System.' Two articles have appeared in journals based on the first study. These documents seem to represent the publicly accessible evidence base presented in support of SDM."*

Response: There is a great deal more information on SDM™ readily available to Mr. Morton. CRC has completed 15 risk assessment studies to date; most of these studies have been published in Summaries of the Annual Risk Roundtables (sponsored by the American Humane Association and the American Public Welfare Association – now APHSA). More on the availability of information on SDM™ is presented below as responses to other allegations.

Morton: *"The first study, funded by the National Center on Child Abuse and Neglect, examined inter-rater reliability and validity of three risk assessment frameworks, SDM, a Washington State model, and a model used by a few counties in California. According to the report, the SDM model attained an inter-rater reliability of .56, higher than the other two models evaluated. Baird and Wagner argue that coefficients between .5 and .6 are generally acceptable (no citation offered by the authors). Literature on inter-rater reliability contradicts this. Generally, inter-*

rater reliability of 80 percent or better is considered acceptable for a widely used instrument (Carmines & Zeller, 1979)."

Response: Mr. Morton has confused a number of different measures here (and in statements that follow). We stated that Cohen's kappa values of .5 to .6 were generally acceptable; this was not a reference to coefficients. Further, we chose two widely used consensus-based models. While Mr. Morton reports that one model was used "by a few counties in California," it was, in fact, a derivative of the widely applied Illinois CANTS model. This is explained in the article.

Morton: *"A recently reported study of another model achieved inter-rater reliability of r=.92 for cumulative risk and r=.96 for overall risk (Leschied, et al 2003). The Texas risk assessment model has achieved an alpha coefficient score of .99 under different conditions. This suggests that it is the proper training of raters, rather than the instrument itself, that determines inter-rater reliability."*

Response: The reliability figures cited by Mr. Morton were based on limited studies conducted during training sessions. (Although Mr. Morton provides no citations, the Texas study we are familiar with was based on a single case and that case was used to train workers to use the risk assessment protocol.) The CRC study was based on 80 cases, 20 each from four different jurisdictions. These cases were rated long after training was conducted. In essence, the ratings Mr. Morton cites were not produced under comparable circumstances. Other studies, conducted outside of training sessions, have produced much lower levels of reliability than those cited by Mr. Morton (see for example Rossi, Shueman, and Budde, 1996). We know of no instance in which the levels of reliability he cites have been achieved in the field.

Furthermore, we presented both percent agreement and Cohen's kappa. These measures represent the accepted method for calculating inter-rater reliability. The symbol "r" represents a correlation coefficient. It is theoretically possible for two raters to never agree on a risk level, yet the correlation between their ratings could be a perfect 1.0. It is, therefore, not a good measure of inter-rater reliability.

In a single paragraph, Mr. Morton cites percentage agreement, Cohen's kappa, and correlation coefficients as if they are comparable measures. The result is a confused, misleading discussion of the issue.

Morton: *"The research design involved raters who were each trained in one of the three models. The report does not identify who did this training or provide information about the experience of the trainers relative to each risk assessment system, simply mentioning an 'expert.'"*

Response: Mr. Morton bothered only to cite a journal article, where space constraints limited the amount of detail that could be presented. Our full research report is on file with OCAN and readily available to Mr. Morton. This report identifies each trainer. Each was an experienced trainer in one system and was selected by a representative of each state's model.

Morton's comment raises two questions: First, why would a federal resource center not have a major study of decision making sponsored by its funding body on file? Second, if it is on file at the resource center, why didn't Mr. Morton read the report prior to sending a critique of our work across the nation?

Morton: *"Baird and Wagner (2000) state, 'In addition, all training sessions included inter-rater reliability testing to ensure that case readers understood the system thoroughly' (p. 846). If raters performed so well in training on the instruments why were the inter-rater reliabilities in the case review so poor? On the other hand, if low inter-rater reliabilities were revealed during the training it would seem to have suggested that the raters were not adequately trained."*

Response: High levels of inter-rater reliability were attained during the training sessions. However, the true test of reliability is the level of consistency obtained in the field. It is important to note, that in an attempt to be completely fair to all models, several different measures of inter-rater reliability were presented and, in every instance, the actuarial instrument had significantly higher reliability than the consensus models.

Morton: *"The authors use the statistic Cohen's kappa to measure reliability. This statistic has been under severe criticism since several authors identified problems with its interpretation (Gwet, 2002a, 2002b, Cicchetti and Feinstein, 1990). The statistic used is known for, 'Its suspicious behavior with respect to the variation of the trait prevalence and to the magnitude of raters' classification probabilities.' (Gwet, 2002b, p. 9)"*

Response: Cohen's kappa and "percent agreement" have been used for 40 years and remain the appropriate tests for measuring inter-rater reliability (G. David Garson, Ph.D., North Carolina State University, Scales and Measures, 2003).

The issued raised by Mr. Morton refers to a circumstance where there are two raters and two choices (e.g., yes/no) and a very high percentage of all ratings fall in one category (e.g., yes). Dr. Gwet argues that the kappa produced in this instance is too low. (We believe that a low

kappa may be deserved given prior probabilities. However, we'll leave this argument to the statisticians involved.) Mr. Morton's point regarding Cohen's kappa is best described as a red herring.

Morton: *"While making an argument for higher reliability, SDM's developers curiously ignore another measure of reliability, Cronbach's alpha, an important measure of the internal reliability of a scale and relevant to a scale's overall validity. According to a CRC staff member, researchers there do not believe that coefficient alpha is relevant to their scales, although other researchers I've asked disagree stating that internal reliability is an important aspect of scale construction."*

Response: Cronbach's alpha is not a proper measure of the reliability (or validity) of risk assessment instruments. Cronbach's alpha measures the extent to which item responses obtained at the same time correlate with each other (G. David Garson, Ph.D., North Carolina State University, Scales and Standard Measures, 2003). This is important when measuring a construct such as depression. Since there is no actual litmus test for depression, it is impossible to correlate an item with observable criterion to validate the item. The next best approach is to hypothesize that all items on a depression scale should have some degree of covariance. Cronbach's alpha is ideal for this purpose.

Maltreatment, on the other hand, is not a construct, but an observable outcome. The relationship between risk factors and maltreatment does not have to be estimated; it can be measured. All risk research stresses the need to avoid covariance and to limit the number of items on a scale. These principles are in direct conflict with maximizing Cronbach's alpha. For risk assessment, it is best when all risk items are totally independent of each other, but each has a

relatively strong relationship to the outcome measure utilized. We don't know who Morton consulted, but no researcher with experience constructing actuarial risk indices should suggest using Cronbach's alpha.

Morton: *"Another statistic missing from the SDM research is the r-value for the scales themselves. This is a measure of the correlation between the scale score and the dependent variable (recurrence of maltreatment). Baird and Wagner (2000) argue that this measure is irrelevant saying, 'But if, as noted above, simple classification is the goal, explained variance in outcomes is of little consequence. What is important is the* degree *(emphasis in original) to which families in different risk groups perform differently. Furthermore, if prediction is not the goal, the issue of false positives and negatives is moot' (p. 851). Do Baird and Wagner make this argument because the explained variance of their scales is actually minimal? Even if the researchers were not convinced of its legitimacy in this context, it would still be appropriate to publish it since it is such a well-accepted standard in studies of this nature."*

Response: Correlation coefficients (or r values) are often reported as measures of the validity of risk assessment instruments. Our article however, discusses at length why conventional measures of validity should be abandoned. Changing the way risk assessment is interpreted and communicated to the public is supported by the National Research Council (Improving Risk Communication, 1989) and two of the leading risk assessment researchers in the mental health field, Monahan and Steadman (1996). Silver and Banks (1998) eloquently summarize the issue: "The best that the risk assessor can do then is to provide the most accurate probabilistic statements of risk possible while maintaining an adequate degree of dispersion among the

probabilistic estimates" (Calibrating the Potency of Violence Risk Classification Models: The Dispersion Index for Risk (DIFR), 1998).

Correlation coefficients are not reported in our paper because they are inadequate measures of validity. To report these measures would continue to focus on the wrong issue and would not assist the child welfare field in any way.

It is particularly ironic that Mr. Morton follows his argument that we should have used Cronbach's alpha with the point that we should have reported the amount of variance explained. He apparently is not aware of the fact that these are contradictory points. We suggest that he pick up a text book on regression and review the need to avoid correlation between factors used to explain variance in a dependent variable.

Morton: *"In reality, simple classification is not the goal. If it ended here, one might accept their premise. However, administrative decisions and actions flow from these classifications. Families with moderate or high-risk levels have cases opened. What is associated with a false positive? A family may be subjected to a non-voluntary intervention. The family may experience coercion to participate in services it may not feel it needs or wants. Considerable government resources may be expended in this effort. Families that do not acquiesce to agency pressure may find themselves in court and under court orders. Deemed uncooperative, they may have their children taken from them simply because they do not comply with agency case plans. Would a family in this situation say false positives are a moot point? A false positive has serious implications for families and the child welfare system as a whole and cannot be dismissed as a 'moot point.'"*

Response: This whole paragraph can only be described as total nonsense. SDM™ does not operate in this manner anywhere. The number of distortions and false assumptions presented by Mr. Morton is truly astounding. For example:

- *"Families with moderate or high-risk levels have cases opened."*
  - No, in every jurisdiction using SDM™ many (if not most) of these cases are not opened for services.
- *"A family (high or moderate risk) may be subjected to non-voluntary intervention."*
  - No, certainly not based on their risk level. Unless the court so orders, non-voluntary intervention is impossible.
- *"Deemed uncooperative, they may have their children taken from them simply because they do not comply with agency case plans."*
  - No, children are removed only when they are found to be unsafe (via a safety assessment), and they remain in placement only if the court agrees they are unsafe.

We suggest that before Mr. Morton again pens such a diatribe he actually check with an agency using SDM™.

Morton: *"Baird and Wagner (2000) argue that classification is, 'a clear indication that such families may require more attention and more services, because cases in this designation tend to "fail" at a higher rate than cases in other classifications.' (p. 852-853)"*

Response: The fact that this sentence makes no sense indicates that something is seriously amiss. This appears to be a deliberate misquote. The correct quote is: "Classification recognizes that a high risk designation (emphasis added) is not a prediction of failure. It is, instead, a clear indication that such families may require more attention and more services because cases in this

designation tend to 'fail' at higher rates than cases in other classifications (pp. 851-852)." Omitting our reference to high risk cases allows Mr. Morton to combine the moderate and high risk categories to make his point. To complete his argument requires further distortion (see next statement).

Morton: *"What threshold of failure constitutes this higher rate that justifies an authoritative intervention by government? The authors suggest that all families that score above one standard deviation below the mean should receive a government intervention (Baird and Wagner, 2000, p. 885). Based on a normal distribution, moderate and high-risk cases would constitute 84.1 percent of substantiated cases, or 62 percent more than are opened currently and more than four times the actual recurrence rate for all cases."*

Response: Our article ends on page 871 (volume 22, numbers 11 and 12, *Children and Youth Services Review*). There is no page 885. The bigger problem is that nowhere in the entire article is there anything that even remotely resembles such a suggestion. Mr. Morton's statement is pure fabrication. Without these two serious distortions, his point that 84 percent of substantiations would be opened for services is totally without foundation.

Morton: *"The child protective services evaluation study in Michigan was based on a non-equivalent comparison group design with post-test only measures of outcomes, the second weakest research design. (Campbell and Stanley, 1966)"*

Response: This is a meaningless (and inaccurate) generalization. Campbell and Stanley were supporters of quasi-experimental research and recognized that the strength of the quasi-experimental design was dependent on the controls used to account for exogenous factors.

In the Michigan evaluation, the comparison and experimental groups were selected from "naturally assembled collectives such as classrooms as similar as availability permits" (Campbell and Stanley, Experimental and Quasi Experimental Designs for Research, 1966, p. 47). Further, these groups were "pre-tested" on a number of variables to measure their equivalence. Most importantly, there were no significant differences in outcomes reported for the two groups prior to implementation of SDM™ in the experimental group. The table presented below demonstrates the level of pre-test equivalence between the comparison and pilot counties.

Table 1

SDM™ and Comparison County CPS Statistics
(17-Month Period Prior to SDM™ Implementation in Pilot Counties)

	Pilot	Comparison
Characteristics at Investigation*		
Prior Abuse/Neglect Substantiation	43%	42%
AFDC Recipients	49%	50%
Foster Care Placement	9%	10%
Opened for Services	88%	88%
Court Involvement	32%	28%
CPS Follow-Up Outcomes at 12 Months		
Re-Referred	20%	18%
Substantiated	9%	10%
Average Number Re-Referrals	.243	.211
Average Number New Substantiations	.097	.106

*Unduplicated count of families substantiated for abuse or neglect during the period. If a family was substantiated more than once, only the first one is represented. CPS follow-up outcomes are observed for 12 months after the first substantiation recorded during the period.

It is extremely doubtful that greater equivalence could be attained from random assignment. True experimental designs, no matter how much they are preferred, are rarely possible in social services. As Carol Wiess notes, "While experimental design has prestige, power, and symmetry, quasi-experimental design often has the overriding virtue of feasibility" (Evaluation Research: Methods of Assessing Program Effectiveness, 1972, p. 73). Furthermore, adherence to rigid research principles does little to inform the field. Noted author, Lisbeth Schorr states, "Evaluators of social programs often put a higher priority on elegant and precise statistical manipulation than providing usable knowledge" (Common Purpose, p. XV).

Michigan should be congratulated for the foresight to evaluate the impact of a policy and procedural change before going statewide with the initiative. Mr. Morton would be well advised to apply the same rigor to system changes that he advocates.

Morton: *"Counties were matched on total population of county; median income of county; percent of population living in poverty; percentage of single parent households; number of substantiated abuse/neglect investigations per month; number of CPS referrals per month; and number of full-time CPS employees, variable with no established relationship to actual case outcomes."*

Response: First, income, poverty, and single-parent households have all been shown to have a relationship to subsequent maltreatment. Second, it was important to match counties by population, CPS referrals, and number of CPS employees to help assure that there were no significant differences in workload between the two groups of counties. Finally, Mr. Morton fails to note that this matching produced a comparison group with no significant differences in case outcome over a 17-month period prior to implementation of SDM™. He also fails to note

that it produced study cohorts with nearly identical risk profiles. The matching scheme produced two very comparable cohorts.

Morton: *"According to the evaluation, lower rates of new substantiations, lower rates of subsequent placements and fewer injuries were recorded in SDM counties. These results are presented in a series of bar graphs displaying percentages, a format that can be greatly deceiving unless one examines the base number. For example, although the substantiation rate for new referrals in cases closed following substantiation is 8.9 percent in comparison counties, and 4.4 percent in SDM counties, when taken as a percentage, this amounts to 15 versus 9 cases respectively. The study does not break the data down by county and reveal how many pilot counties were significantly better than comparison counties. Taken as a whole population of families, differences in one county alone could conceivably account for the differences noted."*

Response: Graphical presentations based on percentage differences are not deceiving. This is the correct way to present differences, especially when different sized cohorts are compared. Mr. Morton's presentation is, however, somewhat deceiving. First, he selectively reports on the smallest sub-sample in the study, closed cases. Second, he ignores differences in the size of the study cohorts and reports raw numbers. There were 206 closed cases in the SDM™ pilot counties and 169 closed cases in the comparison counties. Thus, the SDM™ counties had 40 percent fewer subsequent substantiations despite the fact that the SDM™ cohort was 22 percent larger than the comparison cohort of closed cases. However, we reported this finding in a very conservative format simply indicating the overall rates of subsequent maltreatment for both groups. The fact that SDM™ counties did better on every outcome measure for every subgroup analyzed is a strong indication of the effectiveness of SDM™.

The purpose of combining counties into pilot and comparison groups was to assemble enough cases to produce reliable and stable measures of outcomes. (There were approximately 900 families in each cohort.) Any fluctuations in outcomes found in county-by-county comparisons would be statistically insignificant and more likely a function of small sample sizes than differences in operations.

Morton: *"It is customary for researchers to acknowledge the limitations of their designs and possible alternative hypotheses. The SDM studies curiously do little of this. At the same time, inherent structural limitations of the research designs, questionable statistics, unreported statistics, unanswered questions about the training of raters and the absence of further replication of these studies suggest that while the findings might be encouraging, they are far from definitive in establishing the efficacy of SDM. In fact, there is sound basis for challenging the conclusions being drawn from the findings."*

Response: Limitations and alternative hypotheses are discussed in sections leading to the selection of evaluation methods chosen. As noted earlier, there are no "questionable" or "unreported" statistics. We are appropriately selective in our use of statistics. If Mr. Morton wants to challenge these points, we welcome the chance to publicly debate the issues.

Training was fully explained; Mr. Morton need only read the source documents before making such claims. Further, other studies have been conducted in other jurisdictions by CRC and other researchers, and more studies are underway. A study of the impact of SDM™ on foster care in Michigan has been completed, gone through a peer review, and accepted for publication.

Summary

Mr. Morton's latest publications contain misquotes, misrepresentations, dubious analytical issues, and preposterous assumptions about SDM™ operations. These documents are merely the latest in a long line of misrepresentations. His attacks on the SDM™ model and on the integrity of CRC, the Institute for Human Services, and the people who work in those agencies are readily refuted by facts and published documents. It is our sincere desire that such baseless allegations and arguments be put to rest. Our concern is that wide dissemination of this misinformation, if not countered, will lead administrators unfamiliar with SDM™ and CRC to false conclusions. At a time when the field is searching for tools to help meet federal outcome standards and to secure the safety, permanency, and well-being of children, it is essential that debate over methods and tools continue. But this debate must be undistorted. Rather than continue to write about CRC and SDM™ and send his opinions across the country, I propose that Mr. Morton engage in legitimate professional debate before child welfare professionals and researchers. We are available.

Mr. Morton's proclivity to misquote the work of others is particularly reprehensible. Deliberately misrepresenting the position of others is absolutely unethical. We suggest that Mr. Morton stop these self-serving and unprofessional tactics so we can all turn our attention to the important work of improving the child welfare system.