THE STATE OF NEW HAMPSHIRE
Northern District of Hillsborough County
30 Spring Street
P. O. Box 2143
Nashua, NH 03061
603 669-7410

APR 23 2010

MICHAEL G VALENTINE ESQ
OFFICE OF THE COUNTY ATTORNEY
ONE INDIAN HEAD PLAZA SUITE 3B
NASHUA NH 03060

- 07-E-0236 State of NH v. Thomas Hurley

You are hereby notified that on April 22, 2010, the following order was entered in the above matter

re: DAUBERT HEARING:

(see copy of order attached hereto)

(Abramson, J.)

4/23/2010
Date

/s/ John Safford
Clerk of Court

cc: Marguerite Wageling, Esq.
Mark A. Larsen, Esq.
Devens Hamlen, Esq.
Ross A. McLeod, Esq.
Robin D. Melone, Esq.

THE STATE OF NEW HAMPSHIRE

HILLSBOROUGH, SS.
NORTHERN DISTRICT

SUPERIOR COURT

The State of New Hampshire

v.

Thomas Hurley

No. 07-E-0236

ORDER

Respondent Thomas Hurley ("respondent") is currently being held at the Secure Psychiatric Unit following resolution of the State's petition to commit him for involuntary treatment under RSA chapter 135-E. Regardless, counsel for both parties have requested the Court release its opinion, following a three-day evidentiary hearing. Given the amount of time and effort expended by counsel in the preparation and presentation of evidence, the Court has agreed.

This Court has previously described, in detail, proceedings under RSA chapter 135-E[1], and will not reiterate that description here. See State v. Ploof, Hillsborough County Superior Court, Northern District, No. 07-E-0258 (April 28, 2009) (Order, Abramson, J.) at 1-3.

Prior to trial in Ploof, the Court ruled on the RSA 516:29-a (2007) admissibility of the Static-99, an actuarial tool that, based on the presence or absence of ten static factors, assesses an offender's relative recidivism risk as compared to other sexual offenders. See id. In that Order, the Court provided a

[1] Although RSA chapter 135-E was modified in 2009, those modifications related to timelines established by the statute, and do not impact the Court's RSA 516:29-a analysis of the admissibility of expert testimony in RSA chapter 135-E proceedings.

detailed description of the Static-99, and described changes made to the instrument in 2008 which the Court described as "new norms." Since that time, further changes have been made to the Static-99, and respondent in this case challenges the RSA 516:29-a admissibility of the instrument in its current form (the "Static-99R"). Because many of the underlying factual and legal issues raised in respondent's Daubert motion were addressed in this Court's Order in Ploof, the Court incorporates herein all pertinent rulings not inconsistent with this Order. See id. Indeed, at hearing, the Court noted that it would not reconsider its rulings in Ploof absent a showing that the circumstances giving rise to those rulings had materially changed.

In this case, as in Ploof, respondent was evaluated by a multi-disciplinary team ("MDT") to determine whether he was a candidate for RSA chapter 135-E commitment. Dr. Carol Ball, one of the psychologists on the MDT, testified at the hearing in this matter. As part of its evaluation of respondent, the MDT applied the Static-99. Once the MDT determined what recidivism risk correlated to respondent's score on the Static-99, it considered five additional factors it concluded were relevant to an assessment of respondent's recidivism risk. Based on all of this information, the MDT determined that respondent met the RSA 135-E:2, XII definition of a sexually violent predator.

In January, 2010, Dr. Ball provided an update to the MDT report. This update indicates that although the MDT had considered a factor labeled "intimacy deficits" in evaluating respondent, this factor should be deleted from the report. Dr. Ball scored respondent on the Static-99R, and determined that his score of 5

placed him in the Moderate-High risk category, in contrast to his score of 6 on the Static-99, which placed him in the High risk category. Dr. Ball indicates that despite these changes, she still believes that respondent meets the RSA 135-E:2, XII definition of a sexually violent predator.[2]

As the Court explained in Ploof, the Static-99 was modified in 2008 by the development of "new norms." In October of 2009, the Static-99 was again modified, and the revised instrument is now known as the Static-99R. Essentially, the Static-99R differs from the Static-99 in two important ways. First, the test component of the Static-99 was modified in the Static-99R such that, rather than simply distinguishing between individuals above or below 25 years of age, the Static-99R now differentiates between four age groups, and individuals are given a score from -3 to 1 on this factor, depending on which age group applies. Second, the 2008 "new norms," which were based on two distinct groups of offenders, have been replaced with four offender "sample" types and corresponding recidivism risks. Thus, a clinician applying the Static-99R must now select, based on certain criteria, which of the four sample types a particular offender is most like in order to determine the recidivism risk associated with the offender's score on the test component of the Static-99R.

On October 20, 2009 respondent requested a Daubert[3] hearing, arguing that "[a]ny testimony concerning the Static-99 original (2000) experience tables, or as revised in October 2008 and October 2009 or other mechanistic risk

---

[2] Because Dr. Ball has updated the MDT report to reflect application of the Static-99R, the Court need not address the admissibility of either the original Static-99 or the 2008 "new norms."
[3] Here, as in Ploof, the Court will refer to the relevant hearing as a Daubert hearing, although the hearing may be more properly titled an RSA 516:29-a hearing.

prediction tool," and "[t]estimony from the State's experts concerning predictions of [respondent]'s future sexual dangerousness, should be excluded from trial because it would not comply with the requirements of RSA 516:29-a (2007). Resp.'s Mo. to Exclude, p. 1; see Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 592-94 (1993). As mentioned, the Court held a three-day hearing on the matter. After consideration of the pleadings, the testimony elicited and arguments made at the Daubert hearing, and the applicable law, the Court finds and rules as follows.

At hearing, the Court heard testimony from three expert witnesses about the validity and application of the Static-99, the development, validity, and application of the Static-99R, and the proper method for assessing an offender's recidivism risk. The Court first heard from Dr. Amy Phenix, the State's expert in forensic psychology. Dr. Phenix wrote the protocol for evaluating sexual offenders in California. In addition, she contributes to the website Static-99.org, and is one of three authors of the Static-99R Evaluator's workbook. See State's Ex. 3. Dr. Phenix is involved in training clinicians on the use of the Static-99R, including how clinicians should "match" a particular offender to one of the four sample types in order to ascertain what recidivism risk corresponds to that offender's score on the test component of the Static-99R. Dr. Phenix has testified for both the State and the respondent in sexual offender commitment proceedings.

The Court next heard testimony from Dr. Carol Ball. Dr. Ball testified how she and the other members of the MDT evaluated respondent, and the basis for

and appropriateness of their methodology. She also testified how, subsequent to the MDT's original evaluation of respondent, which was based on the original Static-99, she applied the Static-99R and adjusted the MDT's findings accordingly. This included testimony regarding her understanding of the appropriate method for "matching" an individual to one of the four sample types recognized by the Static-99R.

Lastly, the Court heard from Dr. Brian Abbott, the respondent's expert in forensic psychology. Dr. Abbott has performed more than 1,400 psychological evaluations, for a variety of reasons, during his 25 year career. He has completed between 90 and 95 evaluations in connection with proceedings like RSA chapter 135-E proceedings. In that capacity, he was hired by respondents 85% of the time. However, he has testified on behalf of the State in only one such proceeding. Dr. Abbott indicated that of the 90-95 such evaluations he has conducted, he has found that a given respondent met the statutory definition of sexually violent predator (or other relevant term) 25-30% of the time: 15% were based on a review of the individual's records, in which case Dr. Abbott did not subsequently perform a full evaluation, and 10% were based on full evaluations, in which case Dr. Abbott did not testify at the individual's trial.

Respondent alleges that several pieces of potential evidence do not comply with RSA 516:29-a. Specifically, respondent challenges: the scientific validity of the Static-99R test component (based on the allegedly low inter-rater reliability of the modified age factor), the propriety of the statistical methods used to generate the Static-99R experience tables, and the methodology employed by

the MDT in applying the Static-99R to respondent, in considering and/or using additional factors to adjust any conclusions reached after applying the Static-99R, and in reaching an ultimate conclusion about respondent's recidivism risk. The Court will address each issue, in turn.

The Validity of the Test Component of the Static-99R

As previously noted, the test component of the Static-99R differs slightly from the Static-99, as the method for scoring the age factor has been modified. Before the Static-99R, the age factor only distinguished between individuals above or below 25 years of age: an individual aged 25 or over received a score of 0 on that factor, while an individual aged 18-24 received a score of 1, because empirical evidence indicated that recidivism risk declined as an offender aged. However, as Dr. Phenix testified, a common criticism of the Static-99 was that the age factor did not fully consider the impact of an offender's age because, for example, a thirty-year-old offender would receive the same score on this factor as a sixty-year-old offender, but research indicated that, all else being equal, the sixty-year-old offender's recidivism risk would be lower than that of the thirty-year-old. In an effort to recognize the continued decline in recidivism risk associated with an offender's increasing age, the developers further divided this factor, establishing four age groups with associated scores. Thus, on the Static-99R, the coding form for the age factor appears as follows:

Question Number	Risk Factor	Codes	Score
1	Young	Aged 18 to 34.9 Aged 35 to 39.9 Aged 40 to 59.9 Aged 60 or older	1 0 -1 -3

At hearing, Dr. Ball indicated that respondent falls within the Aged 40 to 59.9 age group, giving him a score of -1 for the age factor in the Static-99R, as compared to a score of 0 on the Static-99. Thus, respondent scored a 5 on the Static-99R, although his score on the original Static-99 was 6.

Although respondent concedes that the age distinctions recognized by the Static-99R better correspond to the declining recidivism rates of aging offenders, he argues that this change nevertheless invalidates the Static-99R because the inter-rater reliability[4] associated with this modified factor is unknown. In support of his argument, he cites a study by *Quesada, Mercado, & Jeglic* which indicated that, on the Static-99, the inter-rater reliability associated with scoring the two-group age factor was the lowest of the ten factors on the test. Now that this factor has been further divided into four age groups, he argues the inter-rater reliability is likely even lower, thus making the Static-99R inadmissible under RSA 516:29-a.

At hearing, Dr. Phenix testified that the inter-rater reliability of the original Static-99 is generally quite good, and that to the extent differences exist, raters are generally within one point of each other in scoring a particular individual. Although she testified that she was aware of the *Quesada* study, she finds its results "incredulous" because a person's age is readily identified, and she

[4] As noted in Ploof, inter-rater reliability measures how likely it is that multiple clinicians or other individuals applying the test would consistently score an individual offender. Inter-rater reliability thus impacts the RSA 516:29-a analysis of a given instrument because it bears on whether testimony "is the product of reliable principles and methods." See RSA 516:29-a, I(b). Although the inter-rater reliability may also bear on the RSA 516:29-a, II(3) question of whether the theories or techniques underlying testimony have "a known or potential rate of error," respondent concedes that inter-rater reliability can be tested, and the Court therefore need not analyze whether the age factor of the test component has, at least, a potential rate of error.

believes that the test is not difficult to apply. For example, she testified that an individual would be considered to be in the 18-34.9 age group until that individual reached his or her 35th birthday, and thus an individual who was 34 years and 364 days old would fall in the 18-34.9 group.

During his testimony, Dr. Abbott expressed concern that the inter-rater reliability of the revised age factor was unknown. Citing the *Quesada* study, he testified that the revision to the age factor increases the likelihood that the inter-rater reliability of the age factor would be relatively low. According to Dr. Abbott's testimony, there are essentially two potential problems associated with low inter-rater reliability: first, that low inter-rater reliability threatens the validity of the group data because the scoring of the Static-99R may not have been consistent across the group, and second, that the individual being compared to the group data may not have been scored in a manner that is consistent with the group.

The Court finds that the inter-rater reliability of the age factor does not render testimony based on the age factor of the Static-99R inadmissible. At this point, the group data for the Static-99R consists of Static-99R scores that were assigned by the test developers: that is, the test developers applied the Static-99R to the members of the "group" in order to generate the Static-99R experience tables.[5] Even if reliable application of the age factor would be challenging for the average clinician, the group data is based on scoring conducted by the very people who designed the age factor and have reached a

[5] With regard to the *Eher, Rettenberger, Schilling, & Pfaffin, 2009* dataset, it is unclear whether the test developers scored the dataset on the Static-99R, or whether they merely checked the inter-rater reliability of other individuals' application of the Static-99R to the dataset. In either case, the inter-rater reliability would be high for the age factor of the Static-99R: "Inter-rater reliability was assessed by having four raters code 27 cases (ICC = .90)." State's Ex. 3, p. 14.

consensus about how it should be applied. Therefore, the Court finds it extremely unlikely that the group data underlying the Static-99R experience tables would suffer from low inter-rater reliability relative to the age factor. As Dr. Phenix testified, it is not difficult to make a determination about whether an individual has or has not reached a particular age, whether that age is 35, 40, or 60. Although the Court acknowledges respondent's argument that the scoring sheet's use of, for example, 34.9 may be confusing to some clinicians, Dr. Phenix credibly testified that the cut-off, as determined by the test developers, is the date at which an individual attains the minimum score for the next age group: his or her 35th, 40th, or 60th birthday. Application of the test component, consistent with this definition of the age cut-offs, does not appear susceptible to low inter-rater reliability, and the only evidence before the Court is that the test developers, in scoring the group underlying the Static-99R experience tables, utilized this method. Accordingly, the Court finds that the inter-rater reliability of the age factor does not threaten the validity of the current group data for the Static-99R.

The remaining issue regarding the age factor is whether the inter-rater reliability indicates that respondent was not scored on the instrument in a manner that is consistent with the scoring of the group data. The Court finds that this concern does not support exclusion of evidence based on the age factor of the Static-99R. Even if the MDT did not score respondent in a manner consistent with the scoring of the group, any error would be readily identified and exposed on cross-examination. Thus, the Court finds that this concern, if applicable, would not serve as a basis for excluding testimony concerning the MDT's scoring

of respondent on the age factor of the Static-99R. See Baker Valley Lumber Co. v. Ingersoll-Rand Co., 148 N.H. 609, 614 (2002) ("[O]bjections to the basis of an expert's opinion go to the weight to be accorded the opinion evidence, and not to its admissibility. The appropriate method of testing the basis of an expert's opinion is by cross-examination of the expert."). Moreover, respondent has not argued that the MDT incorrectly scored respondent on the age factor of the Static-99R.

The Court notes that respondent's challenge to the age factor of the Static-99R focuses on the issue of the inter-rater reliability. Respondent does not claim that the age factor constitutes a new instrument that must be peer reviewed and published prior to being admissible in this case. This may be because respondent has reached the same conclusion as the Court on this issue: that because the age factor was on the original Static-99, and the Static-99R simply altered the scoring of this factor in response to strong statistical support for such a change, the age factor of the Static-99R is not so novel as to render the publication and acceptance of the test component of the original Static-99 inapplicable to the test component of the Static-99R. As respondent concedes, because the relevant data strongly indicate that recidivism declines with age, the empirically-guided revision of this factor poses no bar to the admissibility of the Static-99R. See Resp.'s Mo. to Exclude, p. 11 ("[T]he advent of the revised age item is long overdue based on the large body of literature on the effects of advancing age in reducing sexual recidivism . . ."). Indeed, the theories and techniques underlying the test component of the Static-99R do not differ from

those underlying the Static-99: although the relevant data has led to a difference between the two instruments, the theories and techniques applied to the data were the same.[6] Accordingly, the Court's ruling in Ploof regarding the RSA 516:29-a, II admissibility of the test component of the Static-99 extends to the test component of the Static-99R.[7] Consistent with the foregoing, the Court overrules respondent's RSA 516:29-a challenge to the age factor in the test component of the Static-99R.

In addition to challenging the inter-rater reliability of the age factor of the Static-99R, respondent challenges the entire test component, asserting that the empirical link between each of the ten factors on the Static-99 is too low to statistically support use of the instrument to predict sexual recidivism. Although the method for scoring the age factor was modified on the Static-99R, the ten factors comprising the test component of the Static-99R are the same as those on the Static-99. As the Court outlined in Ploof, the test component of the Static-99 met all of the admissibility requirements of RSA 516:29-a, including general acceptance in the scientific community for use in assessing recidivism risk. See Ploof, p. 9-19. Because the empirical link of the ten factors has not changed, and no new evidence has been submitted on this point, the Court will not revisit its ruling on this issue in Ploof. Accordingly, respondent's objection based on the

---

[6] There is no evidence before the Court that the theories and techniques underlying the test component of the Static-99 and the Static-99R differ. However, the Court is aware that the technique of linear regression used to generate the experience tables for the "new norms" in 2008 and the Static-99R in 2009 was not used with the original Static-99.

[7] Although the Court need not conduct an RSA 516:29-a, II analysis of the test component of the Static-99R, the Court notes that the revision to the age factor has been tested: according to Dr. Abbott's testimony, in revising the age factor, the test developers split the relevant population in half, using one half to develop the revisions and the other half to cross-validate the revised test component.

empirical link between the ten factors comprising the test component of the Static-99R and rate of sexual recidivism is overruled.

In light of the foregoing, the Court finds that the test component of the Static-99R complies with the requirements of RSA 516:29-a. Therefore, expert testimony regarding the test component of the Static-99R, and respondent's score on the test component (on a -3 to 12 scale), is admissible at trial.

The Statistical Methods Used to Generate the Static-99R Experience Tables

Respondent argues that even if his score on the test component of the Static-99R is admissible, expert testimony attempting to qualitatively or quantitatively interpret that score is inadmissible. (For example, testimony associating his test score with a specific probability of recidivism, or testimony indicating that his score places him in a particular percentile relative to other sexual offenders scored on the Static-99R.) Specifically, respondent challenges the admissibility of the "experience tables" generated by the test developers that contain such qualitative or quantitative data. Because the method used to generate the experience tables for the Static-99R is different from the method used to generate the experience tables for the Static-99 and the "new norms," the Court will conduct a thorough review of this issue.[8]

According to Leslie Helmus, a graduate student who worked closely with the test developers during the study and development of the Static-99R,

[8] The Court notes that the actual statistical method of logistic regression used to generate the experience tables is the same method that was used to generate the new norms. Respondent does not argue that use of this statistical method impacts the RSA 516:29-a admissibility of the experience tables. Rather, respondent argues that steps taken prior to applying that statistical method to data, such as identifying four sample types, renders the experience tables inadmissible.

"[a]ctuarial risk scales assess two facets of risk: relative and absolute." Resp.'s Mo. to Exclude, Ex. F, p. 7. Apart from the modification of the age factor, the second way that the Static-99R differs from the Static-99 is that, while the original Static-99 reported a single absolute recidivism risk that corresponded to a given score on the Static-99, and the new norms reported a range of absolute recidivism risk based on two distinct groups of offenders, the Static-99R differentiates between four groups of offenders and reports the absolute recidivism risk of each of the four groups (known as "samples") for any test component score: a "Routine" sample, a "Preselected for Treatment" sample, a "High Risk/High Needs" sample, and a "Non-Routine" sample. See Resp.'s Ex. A, p. 3-4. In other words, for any given score on the test component of the Static-99R, there are four potential absolute recidivism risks, and a clinician must "match" the individual to one of the four samples in order to ascertain the applicable recidivism risk.

Respondent raises several issues regarding the admissibility of the experience tables. First, he argues that the experience tables are invalid because the data used to generate the tables was not subject to random sampling. At hearing, Dr. Abbott explained that random sampling is important because it increases the likelihood that the sample is representative of the larger population. Because the experience tables for the Static-99R were based on whatever data the developers could solicit, Dr. Abbott testified that it was improper to characterize the experience tables as being representative of the overall population of sexual offenders. In response to questioning regarding the

lack of random sampling in the data underlying the Static-99R experience tables, Dr. Phenix opined that, because the Routine sample consisted of every level of sexual offender at the relevant institutions, and was not comprised of a special subset of offender type, this data could be generalized to the overall population of sexual offenders in the same way as a random sample: that is, the data in the routine sample was, in her opinion, as statistically relevant and appropriately collected as a random dataset would have been.

Another criticism of the Static-99R experience tables is the requirement that an individual be "matched" to a sample type in order to determine what probability of recidivism corresponds to that offender's score. Respondent argues that this process necessarily involves clinical judgment, a process which, as noted in Ploof, is only slightly more accurate than pure chance. Order, p. 15. At hearing, Dr. Phenix disputed that this process involved clinical judgment, because a clinician had some guidance in the matching process. Rather, she described the matching process as one of "professional" judgment. In her thesis, Helmus indicates that "[t]he predictive accuracy of structured professional judgment was typically intermediate between actuarial scales and unstructured clinical judgment, although somewhat closer to unstructured judgment." Resp.'s Mo. to Exclude, Ex. F, p. 6.

According to Dr. Phenix, the information used to generate the experience tables for the Static-99R came from a cross-validation of the Static-99. Dr. Phenix testified that four sample types were created because the test developers noted that the base rates of the 29 datasets they were working with varied

considerably: that is, although each of the datasets contained a variety of scores on the test component of the Static-99R, the actual recidivism rates of the offenders were not consistent among the datasets. Dr. Phenix explained that the test developers noted four groupings of dataset recidivism rates. Finding these divisions too important to ignore, the test developers attempted, by way of post-hoc logical inference, to explain the differences between the datasets in each grouping. After looking at the sources of the datasets, the test developers concluded that the differences depended on whether the dataset included a more generalized population of sexual offenders (the Routine sample), a population of sexual offenders who had previously been selected as needing treatment (the Preselected for Treatment sample), or a population of sexual offenders that had previously been identified as being high risk or needing extra correctional attention (the High Risk/High Need sample). The developers noted that the fourth grouping contained populations of sexual offenders that could not properly be considered "routine" but that did not fit neatly in either the High Risk/High Need or the Preselected for Treatment samples. According to Helmus, the Routine sample is "the most representative of the population of all adjudicated sex offenders." Resp.'s Mo. to Exclude, Ex. F, p. 132.

In describing the process of matching an individual to a sample type, Dr. Phenix explained that clinicians should use the Routine sample as the default position, and cannot place an individual in one of the other sample types without justifying such a match. According to Dr. Phenix, the sample types were identified by looking at what characteristics distinguished the sample types from

each other and were common among the datasets within each sample type. For example, the Pre-Selected for Treatment sample is comprised of datasets with a particular characteristic in common: at some point in their history, the individuals in those datasets were selected as needing treatment. Thus, the common characteristic of being selected for treatment distinguished the datasets within this sample type from the other sample types. By contrast, other characteristics of individual datasets may be described in the Static-99R workbook or elsewhere in the literature, but according to Dr. Phenix, these specific descriptors (for example, "sexual homicide perpetrators") do not define a sample type, and simply describe the individuals within a given study or dataset. During her testimony, Dr. Phenix indicated that the matching process involves consideration of only those characteristics that distinguished the four datasets, and does not consider the specific descriptors of a single dataset as a basis for matching an offender to the sample type that corresponds to that dataset. In other words, clinicians should match individuals based on characteristics that are common to the datasets within a sample type and that distinguish those datasets from the other sample types.

During Dr. Ball's testimony, she revealed that she had very recently attended training, by Dr. Phenix, in the matching process. During her testimony, Dr. Ball indicated that, although she would still match respondent to the High Risk/High Needs sample type after her training in the matching process, her justifications for doing so had changed. For example, she testified that one of the important factors was that respondent had "maxed out" – that is, he was held to

his maximum release date. According to Dr. Ball, this was consistent with the sample characteristic of an individual's detention being extended subsequent to a determination that the individual was dangerous. Dr. Ball also testified that, during the matching process, in addition to the relevant characteristics outlined in the Static-99R scoring template, it was appropriate to consider dynamic risk factors that have been empirically linked to recidivism in order to determine which sample type corresponds to the individual. For example, she indicated that it would be appropriate to consider the dynamic risk factors of use of a weapon, or threats of violence, in matching an individual to a sample type.

Dr. Abbott, who heard the testimony of both Dr. Phenix and Dr. Ball prior to taking the stand, opined that the matching process did in fact require the use of clinical judgment. He noted that the unknown inter-rater reliability of this process is a major concern because the recidivism probabilities vary so widely among the four sample types. See State's Ex. 3, p. 5-8. For example, a score of 5 on the Static-99R is associated with a predicted recidivism rate of 11.4% in the Routine sample, but is associated with a predicted recidivism rate of 25.2% in the High Risk/High Needs sample. Moreover, because the matching process is subjective, and the factors considered are not completely clear, he opined that the inter-rater reliability of the process was likely low. Based on these considerations, Dr. Abbott indicated that the matching process amounted to a major change between the Static-99 and the Static-99R, and that the Static-99R should not be used until the inter-rater reliability of the matching process is known.

Based on the foregoing, the Court finds that evidence concerning the High Risk/High Need, Preselected for Treatment, and Non-Routine sample types is inadmissible at trial. Dr. Abbott testified that the data underlying these samples is not statistically relevant or reliable, and the Court heard no testimony refuting that claim. Moreover, Dr. Phenix conceded that the matching process necessarily involves, at best, professional judgment, a process which Helmus describes with regard to predictive accuracy as being closer to clinical judgment than to actuarial assessments. In addition, the method for conducting the matching process remains unclear: Dr. Ball, who had received training on this process within the week preceding her testimony in this case, provided testimony that was inconsistent with Dr. Phenix's description of the relevant considerations. For example, while Dr. Phenix testified that the relevant considerations in the matching process were those factors that were common among the datasets in a particular sample type and that made those datasets different from the datasets in the other sample types, Dr. Ball testified that a variety of factors could be considered in the matching process, such as use of a weapon or threats of violence, so long as those factors were empirically-linked to increased recidivism. Thus, Dr. Ball understood the scope of the relevant considerations to be far broader than that indicated by Dr. Phenix's testimony. Further, Dr. Abbott's criticism regarding the unknown inter-rater reliability of the matching process is well taken: without clear guidelines regarding what factors to consider, or how many "points" or factors must be documented prior to assigning an individual to a particular sample type, the matching process appears subject to inconsistent

application. Finally, although the base rate differences observed by the test developers may be important, the factors identified by the test developers as distinguishing the sample types were identified through post-hoc logical inference, and may not actually be the cause of the differing base rates. Based on the foregoing, the Court concludes that the experience tables associated with the High Risk/High Need, Preselected for Treatment, and Non-Routine sample types are not the product of reliable principles and methods, and are therefore inadmissible under RSA 516:29-a, I(b).

The Court will now address the admissibility of the experience table associated with the Routine sample. Because Dr. Phenix credibly testified that the Routine sample was, for statistical purposes, the functional equivalent of a random sample, the Court finds that Dr. Abbott's criticism on this point does not support exclusion of the Routine experience table at trial. In addition, because the Court has excluded the other sample types, the criticisms and pitfalls of the matching process are inapplicable to the Routine sample. However, respondent raises several arguments regarding the admissibility of the experience tables, generally, which could apply to the Routine sample experience table even if it were the only experience table provided by the Static-99R: (1) because the experience table is based on group data, it cannot be applied to an individual who was not a part of the underlying group; (2) the samples underlying the experience tables do not contain data that is representative of sexual offenders in the United States and/or New Hampshire; (3) that the sensitivity and specificity of the Static-99R is too low to be admissible, and (4) that low base rates have

rendered the Static-99R obsolete because the predictive interval of applying the Static-99R to an individual does not support such use. The Court will briefly summarize each of these arguments.

Respondent's argument regarding group data was also raised in Ploof. Essentially, respondent argues that group data cannot be used to predict the behavior of individuals who are not part of that group. This argument addresses the concern that, because the individual is not part of the group used to generate the data, the applicability of the data to the individual is uncertain. Respondent's argument regarding whether the data used to generate the Static-99R experience tables is representative of United States and/or New Hampshire sexual offenders is simply an extension of this concept: because the datasets used to generate the experience tables contained information on sexual offenders from a variety of institutional and geographic settings, the figures may not accurately represent recidivism probabilities for offenders in the United States in general, or New Hampshire in particular.

Respondents remaining arguments also address the propriety of applying the Static-99R to an individual offender who was not part of the underlying dataset. Respondent argues that the Static-99R is unreliable because, when applied to an individual who is not part of the underlying dataset, the sensitivity and specificity of the instrument are low: that is, the instrument may identify some offenders as "high risk" (or it's functional equivalent in absolute risk terms) who ultimately do not reoffend, and may also identify some offenders as having a low recidivism risk who ultimately do reoffend. Similarly, respondent argues that,

because the overall base rate for sexual recidivism is low, application of the Static-99R to an individual offender results in a predictive interval that ultimately does not support the use of experience table data in assessing the offender's recidivism risk. In that event, Dr. Abbott testified that he would use base rates to assess an offender's risk, rather than experience table data.

All of the foregoing arguments relate to application of the Static-99R experience table data to an individual offender. In Ploof, the Court addressed the group data argument:

> At hearing, Dr. Cobb testified that the Static-99 could not be applied to an individual because the new norms are based on group data, and respondent was not a part of that group. Thus, Dr. Cobb testified, the new norms do not provide respondent's recidivism risk. However, this distinction does not impact the admissibility of the test results. Rather, it indicates only that experts must be careful in describing the new norms. The new norms do not report a particular individual's recidivism risk. The proper characterization of the new norms is that individuals with a particular score on the Static-99 have been found to have a recidivism risk within the range set forth in the new norms. In other words, experts may testify about respondent's score on the one to twelve scale of the Static-99, and that offenders with that score have been found to recidivate at the rate set forth in the ranges of the new norms. See U.S. v. Shields, No. 07-12056-PBS, 2008 WL 544940, at *1 (D. Mass. Feb. 26, 2008) ("[T]he government shall not use the term 'high risk' in reference to a score on a Static-99 test without explaining its meaning in statistical terms (e.g., 'Respondent scored a[n 8] on the Static-99. Individuals with this score have been found to re-offend at [the range set forth in the new norms]. These percentages are group estimates; they do not directly correspond to the recidivism risk of an individual offender.')"). Subject to this clarification, the MDT may testify about respondent's score on the Static-99, and the recidivism range set forth in the new norms that corresponds to that score.

Order, p. 28-29. In this case, a similar distinction in the presentation of evidence can address all of respondent's remaining concerns regarding the Static-99R

Routine experience table: for example, testimony that "in an international study of sexual offenders in a variety of correctional settings, sexual offenders with this score were found to recidivate at (the rate set forth in the Routine experience table)." Such a clarification clearly addresses respondent's argument concerning the applicability of the experience table data to New Hampshire offenders. Moreover, respondent is free to highlight the point on cross-examination by, for example, soliciting testimony that it is unknown whether this data is representative of New Hampshire offenders. See Baker Valley Lumber Co. v. Ingersoll-Rand Co., 148 N.H. 609, 615-6 (2002) ("[O]bjections to the basis of an expert's opinion go to the weight to be accorded the opinion evidence, and not to its admissibility. The appropriate method of testing the basis of an expert's opinion is by cross-examination of the expert."). In addition, because no expert may testify that the experience table necessarily reports respondent's recidivism risk, his concerns regarding sensitivity and specificity, as well as the predictive interval, do not warrant the exclusion of general testimony concerning the experience table. In other words, because no expert may testify that the data in the experience table applies to respondent, or provide testimony that is based on such an application, concerns regarding the sensitivity, specificity, and predictive interval of the instrument as applied to an individual are irrelevant. Based on the foregoing, and subject to the aforementioned limitations, the Court finds that evidence concerning the Static-99R Routine experience table is admissible.

The Static-99R contains two general types of experience tables: those that report absolute risk (as percentage of recidivism probability) and those that report

relative risk. According to Helmus, "effective risk communication should incorporate both absolute and relative risk information." Resp.'s Mo. to Exclude, Ex. F, p. 9. The Static-99R reports relative risk in three ways: Static-99R score percentiles, relative risk ratios, and nominal risk categories. With regard to Static-99R score percentiles, the experience table provides data on the frequency of Static-99R scores in the Routine sample.[9] The table provides two pieces of information for any given score on the Static-99R: it indicates how many individuals (as a percentage of the total population of 4,040 sexual offenders in the sample) received a score on the Static-99R that is lower than or equal to an individual score, and how many individuals received a Static-99R score that is higher than that score. (Thus, if added together, the two figures total 100%).

Based on the evidence, the Court finds that the percentile rankings are admissible in this case. According to Dr. Phenix's testimony, because the percentile rankings were generated based on data from the Routine sample, they are based on reliable data that is as statistically relevant as a random sample would have been. Moreover, the rankings do not assign risk levels, but simply report the observed frequency of each score on the Static-99R in percentile form. Such a report does not amount to a methodology that must be tested, but is simply an alternative way of reporting observed data. Because the Court heard

[9] The percentile rankings were generated based on all available data, from the Routine sample, of Static-99R scores. The Court notes that, for purposes of the Routine sample experience table reporting absolute risk, the dataset included 2,406 individuals, whereas the percentile rankings were based on the Static-99R scores of 4,040 individuals in the Routine sample. Compare State's Ex. 3, p. 13 with Resp.'s Ex. A, p. 2. This is presumably because the test developers had more information regarding Static-99R scores than observed recidivism for individuals in the Routine sample. In any event, the Court heard no testimony indicating that this distinction is relevant for RSA 516:29-a purposes.

no testimony indicating that the calculations underlying the percentile rankings were invalid, the Court finds that this method of reporting the observed rate of Static-99R score in the Routine sample is admissible. In reaching this conclusion, the Court finds that any issues regarding potential jury confusion can be adequately addressed on cross-examination and do not warrant exclusion of this information from trial.

The next category of relative risk data provided by the Static-99R experience tables is relative risk ratio data. This table provides information on the degree to which differences in Static-99R scores impact the likelihood of recidivism. Based on data from 22 datasets regarding 8,047 individuals, the relative risk ratio table provides data on how the risk of recidivism for each score on the Static-99R compares to the risk of an average offender with a score of 2.[10] For example, the table indicates that an individual with a score of 5 is 2.23 times more likely to recidivate than an individual with a score of 2. As the Court has previously noted, Dr. Abbott testified that the data underlying the Static-99R experience tables is not statistically relevant or reliable because it is not the product of random sampling. Although Dr. Phenix testified that the data in the Routine sample is, essentially, the statistical equivalent of a random dataset, the relative risk ratios are based several datasets that are not part of the Routine sample. Accordingly, the Court finds that the relative risk ratio table is the product of an unreliable statistical method, and is therefore inadmissible. See RSA 516:29-a, I(b).

[10] At hearing, there was much debate about whether the relative risk ratios were properly based on a score of 2, because the median score in the relevant population was 3. Because the Court finds that the relative risk ratio table is inadmissible on other grounds, it need not reach this issue.

The last form of relative risk data provided by the Static-99R consists of nominal risk categories. Essentially, the test developers have assigned qualitative words like low, moderate, or high to various scores on the Static-99R. Rather than adjusting all of the nominal risk categories in developing the Static-99R, the test developers simply expanded the "low" category to include Static-99R scores of -3, -2, and -1 in addition to the scores of 0 and 1. In her thesis, Helmus notes that "[n]ominal risk categories are interpreted inconsistently . . . and with more errors than numerical information." Resp.'s Mo. to Exclude, Ex. F, p. 9. Moreover, interpreting nominal risk phrases like

> "moderate risk sex offender" requires knowledge of the risk posed by all sex offenders. If the base rate for sexual recidivism is 70%, then the release of a "moderate" risk sex offender invokes concern. If the base rate is 5%, however, then the release of a "moderate" risk sex offender is much less problematic.

Id. at 8. In concluding her thesis, Helmus noted that "[f]urther research is [] needed . . . to assess whether the cut-off scores for the relative risk categories for Static-99 (i.e., low, moderate-low, moderate-high, and high) should be revised for Static-99R." Id. at 133.

Based on the foregoing, the Court finds that the nominal risk categories are inadmissible in this case. Because these categories are interpreted inconsistently, they run the risk of confusing, rather than aiding, the jury.[11] This conclusion is bolstered by the fact that, at hearing, the experts agreed that local base rate information for sexual offenders in New Hampshire is currently

[11] Although the rules of evidence do not apply in this proceeding, see RSA 135-E:10, the Court nevertheless retains discretion over the admissibility of evidence. See State v. Lopez, 156 N.H. 416, 420 (2007) ("The admissibility of evidence is generally within the trial court's sound discretion").

unavailable. Thus, it is difficult to interpret the nominal risk categories in a meaningful way relative to New Hampshire sexual offenders. In addition, Helmus has indicated that the relevant cutoff scores for the Static-99R may need to be revised. Based on the evidence before the Court, evidence regarding the nominal risk categories, at best, would be confusing to a jury, and, at worst, would amount to unreliable evidence. See RSA 516:29-a, I(b). Accordingly, the Court finds that the nominal risk categories are inadmissible.

The Evaluation By and Conclusions Of the MDT

Based on the evidence presented at the Daubert hearing, the Court finds that the evaluation by and conclusions of the MDT partially meet the requirements of RSA 516:29-a. There are three main components of this evidence: the MDT's application of the Static-99R with respect to respondent, the other factors the MDT considered in determining respondent's recidivism risk, and the MDT's final conclusion about respondent's recidivism risk. The Court has previously ruled that evidence concerning the test component of the Static-99R, the Routine experience table, and the Static-99R score percentile rankings is admissible in this case. Because respondent does not argue that the MDT erred in its technical application of the Static-99R to respondent, the Court finds that evidence regarding the MDT's application of the Static-99R to respondent is admissible in this case. The Court will now address the remaining issues concerning the evaluation by and conclusions of the MDT.

In addition to scoring respondent on the Static-99R, the MDT also considered several additional factors in determining whether respondent met the

statutory definition of sexually violent predator. Although the MDT originally considered five such factors, Dr. Ball's update indicates that the "intimacy deficits" factor should be deleted from the MDT's report. At hearing, Dr. Ball testified that she still viewed this factor as relevant, but deleted it from the report because she knew the propriety of considering this factor was debatable. Because Dr. Ball has deleted this factor from the MDT's report, the Court will not address the RSA 516:29-a admissibility of testimony regarding this factor, and will only address the admissibility of the four factors currently contained in the MDT's report (as updated by Dr. Ball).

The four factors considered by the MDT were: (1) sexual deviancy; (2) sexual preoccupation; (3) anti-social orientation; and (4) attitudes supportive of sexual assault. State's Ex. 2, p. 7-9. At hearing, Dr. Ball testified that each of these factors have been empirically linked to increased sexual recidivism. Although Dr. Abbott conceded that the factors were empirically linked to increased sexual recidivism in the study underlying the Stable 2007, he noted that the populations used to establish the statistical link contained sexual offenders who had been released into the community under either probationary or parole system supervision. Thus, Dr. Abbott questioned whether it is appropriate to rely on this statistical link as a basis for considering these factors in relation to an individual who is presently incarcerated. Like the "group data" criticism of the experience tables, this issue may be properly addressed by careful attention to the manner in which the evidence is presented. For example, "Respondent has a history that is consistent with (a particular factor). This factor

has been statistically linked with increased recidivism in sexual offenders who have been released into the community on probation or parole. However, it is unknown whether this factor would show a similar link to increased recidivism in individuals who have not been released into the community." Consistent with this limitation, the Court finds that evidence of the additional factors considered by the MDT need not be excluded on the basis that their empirical link to increased sexual recidivism may not apply to incarcerated sexual offenders.

In addition to questioning the relevance of the statistical link between these factors and increased recidivism in sexual offenders released into the community under probationary or parole system supervision, respondent raises individual challenges to each of the factors considered by the MDT. The Court will address each factor, in turn.

The first additional factor considered by the MDT is that of "sexual deviancy." State's Ex. 2, p. 7. According to the MDT report, respondent exhibits this characteristic in several ways: he has previously indicated that he cannot control his sexual urges, he has admitted to grooming young males for sexual encounters, he attempted to place himself in positions where he would be likely to encounter "desirable" targets, and he acted out sexually during treatment. Respondent argues that this factor should be excluded because it amounts to double counting. Id. According to respondent, sexual deviancy includes several categories that are addressed on the Static-99R, including prior sex offenses, index nonsexual offense, any stranger victims, any unrelated victims, and any male victims. None of these categories is mentioned in the MDT as a justification

for finding that respondent exhibits sexual deviancy. Although the MDT report notes the grooming of young-looking males, a fair reading of the report reveals that the relevant concern is the act of grooming, not the gender of the anticipated target. Based on the foregoing, the Court finds that evidence concerning respondent's characteristics that are consistent with the factor of "sexual deviancy" is admissible. However, the scope of admissible testimony concerning this factor is limited to characteristics that are not already addressed by the test component of the Static-99R.

The next factor considered by the MDT is "Sexual Preoccupation." State's Ex. 2, p. 7. According to the MDT report, respondent exhibits this characteristic because he has expressed an inability to control his sexual impulses and consequently requested chemical treatment on two prior occasions. Id. Respondent's sole argument regarding this factor involves its empirical link to increased sexual recidivism in incarcerated individuals.[12] Because the Court has already addressed this argument, the Court finds that testimony concerning this factor is admissible. However, the scope of admissible testimony concerning this factor is limited to characteristics that are not already addressed by the test component of the Static-99R.

The third factor considered by the MDT is "Anti-Social Orientation." State's Ex. 2, p. 7-8. The MDT report indicates that respondent exhibits this

[12] Although respondent argues in his Motion that the MDT provided no basis for considering sexual preoccupation as a risk-enhancing factor, Dr. Abbott conceded at hearing that this factor was empirically validated in the study associated with the Stable 2007. Moreover, a quick review of the Sexual Offender Risk and Need Assessment Framework website (http://soraf.cyzap.net/zap_site/docs/zaps-mr-tab1-85.htm?Stable-2007%26copy%3B-assessment) reveals that sexual preoccupation is addressed on the Stable 2007.

characteristic in several ways: he has a lengthy history of inappropriate sexual behavior, including conduct as a juvenile; his disciplinary record in the prison system reflects both impulsivity and a lack of general behavioral control; and he has a limited and unstable work history. Id. Respondent argues that testimony concerning this factor is inadmissible because the Static-99R already addresses prior sentencing dates and prior sexual offenses. The Court agrees that some of the MDT's considerations regarding the applicability of this factor to respondent are addressed by the Static-99R, and testimony concerning those characteristics in connection with this factor would therefore be inadmissible. However, some of the MDT's considerations regarding this factor do not involve prior sentencing dates or prior sexual offenses (for example, respondent's work history or nonsexual disciplinary infractions). Based on the foregoing, the Court finds that some testimony concerning this factor is admissible. The scope of admissible testimony concerning this factor is limited to characteristics that are not already addressed by the test component of the Static-99R.

The final factor considered by the MDT is that respondent exhibits "Attitudes Supportive of Sexual Assault." State's Ex. 2, p. 8. The MDT report indicates that respondent exhibits this characteristic in several ways: he references sexual exploits at every opportunity; he has previously solicited and possessed publications from the North American Man-Boy Love Association; in 1993, he drafted a letter to The Correspondence Foundation in Athens, Greece inquiring as to the "sex laws with boys" in Greece and requesting photos and/or correspondence from young males; and he filed a federal complaint after prison

officials prohibited him from possessing a publication that advocates adult-child sexual relations. Id., p. 8-9. Respondent's sole argument regarding this factor involves its empirical link to increased sexual recidivism in incarcerated individuals. Because the Court has already addressed this argument, the Court finds that testimony concerning this factor is admissible. However, the scope of admissible testimony concerning this factor is limited to characteristics that are not already addressed by the test component of the Static-99R.

The Court will now address the admissibility of the MDT's final conclusions regarding respondent's recidivism risk. In her January 2010 update, Dr. Ball reached her final conclusions about respondent's recidivism risk by assessing respondent using the Static-99R High Risk/High Need experience tables, and then considering how the aforementioned additional four factors impact respondent's recidivism risk. Because the Court has previously determined that the High Risk/High Need experience table does not comply with RSA 516:29-a, Dr. Ball may not testify about the recidivism rates contained in those experience tables, nor may she provide testimony that is based on those recidivism risk probabilities.

Furthermore, testimony by Dr. Ball regarding how respondent's individual recidivism risk is impacted by the presence of sexual deviancy, sexual preoccupation, anti-social orientation, and/or attitudes supportive of sexual assault would necessarily be based on clinical and/or professional judgment. As previously noted, clinical judgment is unreliable, and professional judgment is more like clinical judgment than actuarial assessment. Moreover, respondent

cites a study by Hanson indicating that consideration of additional factors outside of the Static-99 may reduce the overall predictive accuracy of an assessment. See Resp.'s Mo. to Exclude, p. 29. Accordingly, the members of the MDT may not testify that respondent's individual recidivism risk is necessarily impacted by sexual deviancy, sexual preoccupation, anti-social orientation, and/or attitudes supportive of sexual assault, and may only testify generally that these factors been linked to increased recidivism rates in sexual offenders who have been released into the community under probationary or parole system supervision. See RSA 516:29-a, I(b) (requiring that testimony be the product of reliable principles and methods).

In addition, as in Ploof, the Court has heard no testimony indicating that a clinician may predict a particular offender's recidivism risk based on a single actuarial instrument. See Order, p. 13. Accordingly, the members of the MDT may not testify about respondent's recidivism risk based solely on his Static-99R score. See RSA 516:29-a, I(b) (requiring that testimony be the product of reliable principles and methods). The State may introduce evidence of respondent's score on the Static-99R, and the recidivism risks set forth on the Static-99R Routine sample experience tables, but may not testify that respondent's recidivism risk is necessarily consistent with that data. Moreover, the State may not introduce testimony indicating that, in light of the additional factors considered by the MDT, respondent's recidivism risk is necessarily higher than that reported by the Routine sample experience tables.[13]

[13] This is because, as the Court has previously noted, the empirical link between increased sexual recidivism and the additional factors considered by the MDT was based on a statistical

In considering the evidence presented at hearing and arriving at the above conclusions, the Court credits the testimony of the experts relatively equally. Dr. Phenix has a greater basis than Dr. Abbott for understanding the development and application of the Static-99R. However, although the Court found Dr. Phenix's testimony to be generally straightforward and credible, her deep professional connection to the Static-99R is not without the potential for creating bias in favor of the validity and admissibility of the instrument. Similarly, Dr. Abbott's professional record of involvement in sexually violent predator cases evinces a potential bias against the State: out of 90-95 sexually violent predator evaluations, Dr. Abbott has testified on behalf of the State only one time. Nevertheless, like Dr. Phenix, the Court found Dr. Abbott's testimony to be, generally, straightforward and credible. However, the Court found the importance of many of Dr. Abbott's valid criticisms of the Static-99R to be somewhat overstated, insofar as he asserted them as bases for excluding the Static-99R rather than simply being worthy of mention on cross-examination of a testifying witness. Because the Court found both experts to be generally credible, the Court gave their testimony full consideration in reaching its conclusions in this case.

Respondent argues in his pleadings that the "more searching inquiry" involved in State v. Cressey applies to the evidence at issue in this case. See

---

correlation between those factors and the rate of sexual recidivism among sexual offenders in the community but under either probationary or parole system supervision. Although the empirical link may well extend to sexual offenders released into the community without any supervision, that link has not yet been established. Therefore, in this case, the jury must be afforded an opportunity to give the evidence regarding the established empirical link whatever weight the jury sees fit.

cannot provide testimony regarding this subpart of the statutory definition of sexually violent predator, experts cannot provide an opinion as to whether respondent meets the overall statutory definition. They may, however, provide testimony regarding whether respondent meets other subparts of the definition, see RSA 135-E:2, XII, so long as such testimony would not be inconsistent with this Order.

Summary of Permissible Testimony Under RSA 516:29-a.

In sum, the Court finds that the following **expert** testimony is admissible at the trial in this matter: testimony regarding the MDT's use of the Static-99R to assess respondent's recidivism risk, and his score (on a -3 to 12 scale) on that test; testimony regarding the recidivism risk, as set forth in the Routine sample experience table, that corresponds to respondent's score on the Static-99R; testimony regarding the Static-99R score percentile rankings that were generated based on data from the Routine sample; and testimony about the empirically-validated link between sexual deviancy, sexual preoccupation, anti-social orientation, and/or attitudes supportive of sexual assault to increased sexual recidivism in offenders released into the community under probationary or parole supervision, and testimony that respondent displays characteristics that are consistent with those factors (subject to the limitation that any characteristics addressed by the Static-99R may not be discussed in connection with these additional factors).

In addition, **factual** testimony regarding the MDT's justifications for finding the presence of the relevant additional factors is also admissible at the trial in this

matter, as well as factual testimony supporting the MDT's scoring of respondent on the Static-99R. This is not an exhaustive list of permissible factual testimony.

Pursuant to RSA 516:29-a, the Court will not permit the following forms of **expert** testimony: testimony that the jury should give a particular amount of weight to respondent's sexual deviancy, sexual preoccupation, anti-social orientation, and/or attitudes supportive of sexual assault; testimony attempting to quantify the impact of those factors on respondent's recidivism risk; testimony that respondent's recidivism risk is necessarily lower than, higher than, or equal to the recidivism risk in the Routine experience table associated with his score on the Static-99R; testimony regarding the nominal and relative risk data set forth in the Static-99R experience tables; testimony attempting to assign respondent either a quantified (for example, percentage) or a qualified (for example, words like high, medium or low) recidivism risk; and testimony opining whether respondent meets the statutory definition of sexually violent predator.

SO ORDERED.

4/22/10
Date

Gillian L. Abramson
Presiding Justice