4 STATISTICAL TERMINOLOGY

Authors of reports and studies on risk assessment tools adopt a range of descriptions of statistical outcomes. For the sake of consistency in the tables below, the RMA has adopted the following descriptors drawing from the most prevalently used terminology in the literature.

There is a useful explanation that uses an analogy of the bull's-eye on a dartboard: actually hitting the bull's-eye represents accuracy; landing shots together indicates good reliability. Considering this, hitting the bull's-eye and landing all the shots together would convey both accuracy and precision (Viera and Garrett, 2005).

Statistical Analyses

	Definition	Interpretation
Brier Scores	The Brier score is defined as the average quadratic difference between the predicted probability and the binary outcome. Its purpose is to measure the 'calibration' of a set of probabilistic predictions.	The Brier score ranges from 0 to 1. The best possible score is 0, indicating total accuracy. The lowest possible score is 1, meaning that the forecast was wholly inadequate.
Intra Class Correlation Coefficient (ICC) – descriptors; poor/moderate/excellent	An ICC score represents the estimation of the correlation between two scores. It measures the magnitude of agreement for inter-rater reliability.	ICC values range from 0 to 1 and are typically reported with two decimal points, e.g. .75. Cicchetti (1994) recommends the following thresholds: • <.40 = 'poor' • .40 to .75 = 'moderate'

		• .75 to 1.0 = 'excellent'
Kappa (.) Coefficient – descriptors; poor/average/excellent	The Kappa Coefficient (.) measures the agreement between two individuals.	A Kappa is always less than or equal to a value of '1.' A value of '1' implies perfect agreement and a value of '-1' implies perfect disagreement. Recommended thresholds are: • < 0 Less than chance agreement • 0.01– 0.20 poor agreement • 0.21– 0.60 average agreement • 0.61– 0.99 excellent agreement
Effect sizes	The effect size quantifies the size of the difference between two groups. It is calculated as the standardised mean difference between the two groups: mean of Group A minus mean of Group B; the total of which is divided	Effect sizes can be interpreted in terms of the percentiles or ranks at which two distributions overlap. Interpretations of effect sizes are dependent on the assumptions that the two groups are

	by the standard deviation.	normally distributed and have the same standard deviations.
Pearson Correlation Coefficient– descriptors; small/moderate/large	Pearson correlation coefficient measures the association between a predictor variable and the outcome.	The values of r can range from '-1' to '1,' with '0' indicating that there is no relationship between the predictor variable and the outcome. Positive values indicate that the high scores are associated with increased recidivism; whereas negative values indicate that high scores are associated with decreased recidivism. According to Cohen (1988), r values may be interpreted as follows: .10 are small, .25 are moderate, .40 are large.
Sensitivity and Specificity	Sensitivity is the ability of a test to correctly classify an individual as possessing a particular	The higher the value of sensitivity, the greater the ability of the measure being

	characteristic (e.g. offending). Specificity is the ability of a test to classify an individual as *not* possessing a particular characteristic (e.g. not offending). Sensitivity is inversely proportional with specificity, meaning that as the sensitivity increases, the specificity decreases.	tested to correctly identify individuals. For instance, a sensitivity of 62% on a risk assessment tool indicates that it has the ability to correctly classify just under two-thirds of individuals who will reoffend. For specificity, higher values indicate the ability of a measure to correctly identify who will *not* possess certain characteristics (e.g. who will not go on to reoffend).
Z+ - descriptors; small/moderate/large	Measures the association between the predictor variable and the outcome. These two groups usually comprise of the (1) recidivists and (2) the non-recidivists, separated by the difference in their scores obtained on risk assessments.	According to Cohen (1988), d values may be interpreted as follows: .20 is considered 'small'; .50 is considered 'moderate'; .80 is considered 'large'.
Receiver Operating Characteristic (ROC) Curve – descriptors; low/moderate/high	In the context of risk assessment, the ROC curve is a plot that shows the probability	Several different indicators can be calculated from ROC curves. The

	that a measure will correctly identify persons as recidivists or non recidivists (Mossman, 1994; Rice and Harris, 1995). It is a plot of the 'hits' (the proportion of recidivists correctly identified as recidivists) against the 'false alarms' (the proportion of non-recidivists identified as recidivists).	most commonly used indicator is the 'area under the curve' (AUC) (see below).
Area Under the Curve (AUC) – descriptors; low/moderate/high	The area under the curve (AUC) for the ROC curve is a useful summary statistic for the extent to which a measure discriminates between recidivists and non recidivists.	AUC values can range from '0' to '1'. They can be interpreted as the probability that a randomly-selected recidivist has a worse score than a randomly-selected non-recidivist. Values between '.51' and '1.0' indicate positive associations with recidivism; whilst values between '0' and '.50' indicate that predictions are no better than chance.

		Using Cohen's (1988) d values as a guide, AUC values may be interpreted as follows: .556 is considered 'low'; .639 is considered 'moderate'; .714 is considered 'high.'