# Appendix 3.D: Description of Agreement Statistics

Agreement between two raters (e.g., between two human readers or between the final human scores of record and the scores produced by an Automated Scoring system) is measured using the following statistics:

* Standardized mean difference (SMD)
* Kappa (Cohen, 1960) and quadratic weighted kappa (Cohen, 1968)
* Pearson correlation
* Percentage of exact and adjacent agreement

These statistics target different aspects of rater agreement. Consider as an example these statistics computed with the final human scores of record as reference scores and the scored produced by an Automated Scoring system as focal scores.

* The standardized mean difference (the difference between the mean of the focal scores and the mean of the reference scores divided by the pooled standard deviation) measures the bias in the scores assigned by the Automated Scoring system. If SMD is positive, the system tends to assign higher scores than the final human scores. Conversely, if SMD is negative, the systems tend to assign lower scores.
* Kappa and quadratic weighted kappa (QWK) measure the agreement between two raters over the agreement expected by chance. Generally, kappa and QWK range between 0 (no agreement) to 1 (perfect agreement), although smaller and larger values are technically possible. The difference between kappa and QWK is that kappa penalizes all disagreements equally, whereas QWK penalizes larger differences more; in fact, the penalty is the square of the difference. (The precise mathematical formulation of kappa and QWK is too complex for this Appendix.)
* Pearson correlation measures the linear correlation between the ratings. Its values range from -1 (perfect negative agreement) to 1 (perfect agreement) with a correlation of 0 showing no agreement. In contrast to SMD, it does not take possible bias into account. In contrast to kappa and QWK, it does not take the agreement expected by chance into account.
* Finally, exact agreement is the percentage of scores for which both raters agreed exactly. Adjacent agreement is the percentage of scores for which the ratings by the raters are within one score point.