Instructions:

Agreement of scores and score interpretation between assessors who use the rubric is an important goal that must be met before data can be scored. Therefore, before being exposed to the official data used in any research study, novice assessors go through a fairly extensive training process.

A sample training process:

Novice assessors go through 8 physics problems and score at least 5 students' written solutions using the rubric. They should compare their scores with an expert assessor and discuss their results for every problem. They can adjust their score after the discussion. Usually after discussion, the agreement is pretty high. Agreement within 1 (each category of the rubric is scored from 0-5) could go up to 100% after discussion. When achieving exact agreement between raters 60% and the agreement within 1 90% before discussion, the training can be stopped. The training process can be repeated (using different sets of students' written solutions) if necessary.

In addition to percent agreement, another statistical measure of reliability can also be used. Kappa (Cohen, 1960; Howell, 2002) is a measure of raters' exact score agreement after correcting for expected agreement by chance. Weighted kappa is an extension of the kappa measure that considers the degree of difference in raters' scores (Cohen, 1968). Scores that are closer (such as agreement within one score) are given more weight in calculating the kappa agreement score than scores which differ more substantially. A kappa value above 0.60 is considered by some researchers to indicate "substantial agreement" and a value above 0.80 is considered "almost perfect agreement" (Landis & Koch, 1977). You may refer to Jen Docktor's PhD dissertation (Docktor, 2009) for more information.

Reference:

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 10, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213-220.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Howell, D.C. (2002). *Statistical methods for psychology (5th ed.)*. Pacific Grove, CA: Thomson Learning, Inc.

Docktor, J. (2009). Development and Validation of a Physics Problem-Solving Assessment Rubric. Doctoral dissertation, University of Minnesota.