# CRASE® Essay Scoring Model Performance Based on Proof-of-Concept and Operational Engine Trainings

**Scott W. Wood**

Since the creation of CRASE® in 2007, the CRASE Research team has trained the engine on hundreds of constructed-response items and essay prompts. These scoring models have been used in formative, summative, and research assessments across a variety of content areas—English language arts, mathematics, science, government—and for grade levels ranging from 4th to 12th grade. Most models have been **prompt-specific**, meaning that hand-scored response data for a single item were used to train the model, though ACT also has experience working with **cross-prompt** scoring models—that is, models trained on multiple similar essay prompts simultaneously.

This Data Byte presents an overview of the human-CRASE scoring metrics that ACT observed between 2016 and 2019 on writing prompts. The next section describes the two evaluation metrics considered in this paper. This is followed by a summary of CRASE essay scoring performance. The Data Byte concludes with commentary about factors that may affect essay scoring accuracy.

## Automated Scoring Metrics

After training a scoring model in CRASE, a blind-validation sample (i.e., a sample of responses not used for training) will be scored. This sample is used to calculate automated scoring metrics that estimate how the scoring model would be expected to perform on new responses during operational scoring. This research focuses on human-CRASE metrics evaluated **relative** to corresponding human-human metrics from hand scoring.

Quadratic weighted kappa (QWK) is a popular metric used in the automated scoring literature. QWK is a measure of rater agreement that penalizes for disagreement between two independent raters. Small differences between the raters lead to small penalties; larger differences lead to larger penalties. QWK can be thought of as a slight variation of a Pearson correlation, as the two metrics will often be within .01 of each other in practice.

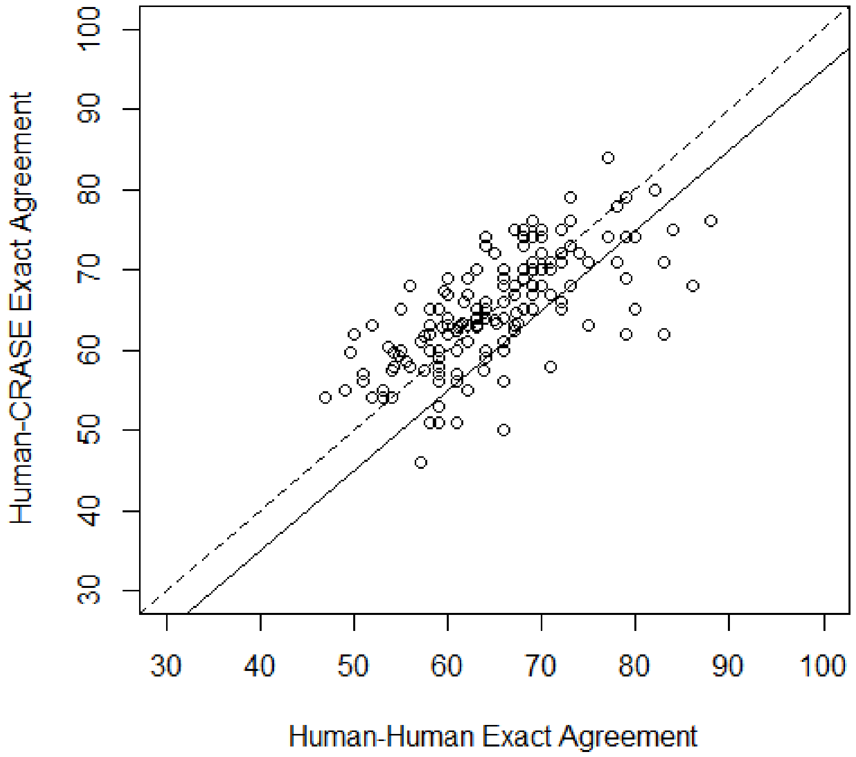
If two independent human raters obtain a certain QWK, the goal is to achieve a similar human-CRASE QWK. Industry-standard criteria suggest that, if the human-CRASE QWK is less than the human-human QWK, then the human-CRASE QWK should be within .10 of the human-human QWK (Williamson et al., 2012).

Some stakeholders prefer the exact agreement rate due to its simplicity. The exact agreement rate is the percentage of essays for which two independent raters assign the same score. The goal is to achieve a human-CRASE exact agreement rate similar to the human-human exact agreement rate. Industry-standard criteria suggest that, if the human-CRASE exact agreement rate is less than the human-human exact agreement rate, then the human-CRASE exact agreement rate should be within 5.125 percentage points of the human-human exact agreement rate (McGraw-Hill Education CTB, 2014).

## CRASE and Essay Scoring

Between 2016 and 2019, the CRASE Research team trained models for 253 essay prompt-dimension combinations across multiple assessment programs. (For prompts scored on multiple dimensions, an independent model is produced for each dimension.) Of these 253 prompt-dimension combinations, human-human scoring metrics were available for 165 (the remaining combinations included only a single human rater). The human-human and human-CRASE exact agreement rates for these 165 prompt-dimension combinations are plotted in Figure 1.

**Figure 1.** Human-Human and Human-CRASE Exact Agreement Rates, by Prompt-Dimension

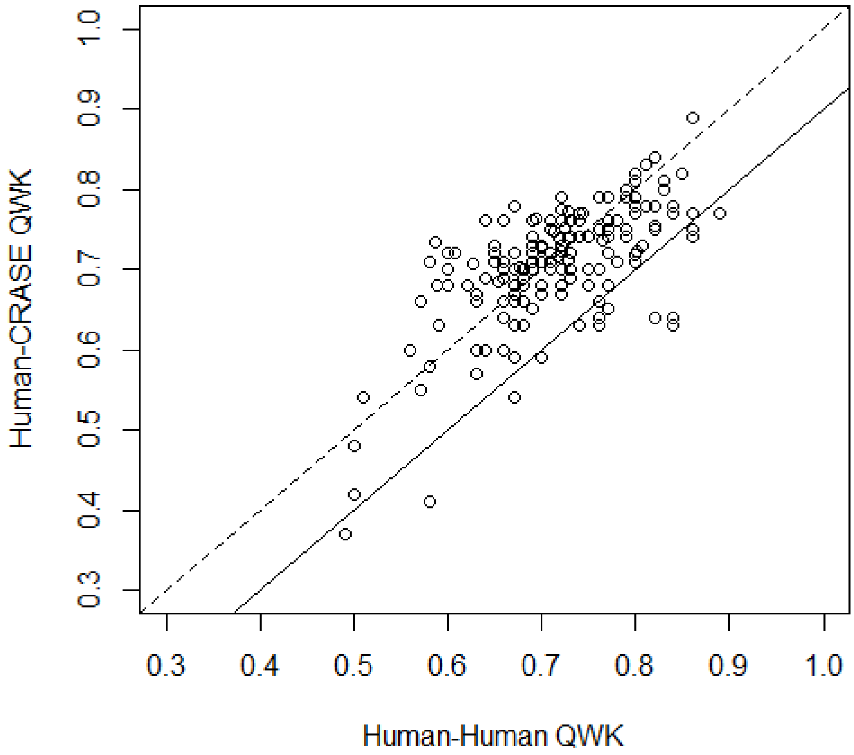


Note that there are two reference lines in the plot. The dashed line marks where the human-human exact agreement rates are equal to the human-CRASE exact agreement rates. The solid line marks where the human-CRASE exact agreement rates are 5.125 percentage points less than the human-human exact agreement rates. To meet industry standards regarding exact agreement rate, a prompt-dimension point must be on or above the solid line.

Across the prompt-dimension combinations, 101 were on or above the dashed line, 140 were on or above the solid line, and 25 were below the solid line. This means that 85% of the prompt-dimension combinations met or exceeded industry standards for exact agreement rate.

Figure 2 shows the human-human and human-CRASE QWKs for the 165 prompt-dimensions combinations for which human-human QWKs are available. Once again, two reference lines are included. The dashed line marks where the human-human QWKs are equal to the human-CRASE QWKs. The solid line marks where the human-CRASE QWKs are .10 less than the human-human QWKs. To meet industry standards regarding QWK, a prompt-dimension point must be on or above the solid line.

**Figure 2.** Human-Human and Human-CRASE Quadratic Weighted Kappa, by Prompt-Dimension



Across the prompt-dimension combinations, 90 were on or above the dashed line, 151 were on or above the solid line, and 14 were below the solid line. This means that 92% of the prompt-dimension combinations met or exceeded industry standards for QWK.

## Factors that Affect Scoring Accuracy

Between 2016 and 2019, the CRASE engine performed well, scoring essay prompts reliably and accurately, with 85% of prompt-dimension models meeting benchmarks for exact agreement rate and 92% meeting benchmarks for QWK. This was due, in part, to the feature set developed for CRASE, which was assembled using psychometric and English language arts expertise. The feature set was designed to apply to a variety of writing rubrics across customers and grade levels.

If human-CRASE metrics do not meet industry-standard benchmarks, we may need to try different parameters and settings during engine training. If engine performance continues to be troublesome for a prompt-dimension combination, there are several key questions to consider when determining if an essay prompt is suitable for automated scoring:

> **Does the hand scoring rubric give clear guidance on what distinguishes essays of different score points?** Unclear rubrics lead to hand scoring data with additional noise, which makes it difficult to train a reliable model.
>
> **When conducting hand scoring, were best practices followed to ensure that the hand scoring data were of the highest quality?** Best practices include appropriate training, the use of qualifying tests, the periodic use of validity papers to catch rater drift, and score monitoring by expert readers. Failure to use these best practices leads to hand scoring data with additional noise, which makes it difficult to train a reliable model.
>
> **Are two or more hand scores available for each prompt-dimension?** Although the CRASE team can train an engine with only one human score per prompt-dimension, having two or more scores allows for the calculation of human-human metrics, yielding a more complete picture of how accurately the CRASE engine performs.
>
> **Do the hand scoring data have adequate coverage at all score points?** Often, scoring data will include few or no responses at the lowest and/or highest score points. A lack of data at the extreme score points hinders the scoring models from accurately identifying these kinds of essays operationally.
>
> **Are there at least 100–200 essays at each score point?** As with all statistical modeling, the scoring model will be more precise and reliable if it is built with more data. From our experience, having at least 100–200 essays per score point is enough to produce a reliable scoring model.

The data presented in this Data Byte show that CRASE+ can score most essay prompts and dimensions with acceptable accuracy. Additional findings from 2022 and 2023 show a similar pattern. For example, during a 2022 proof-of-concept study using essays from the ACT® writing test, CRASE was able to achieve exact agreement rates and QWKs that met or exceeded industry standards across all four domains of writing on the ACT writing test rubric. This evidence, and the evidence in this Data Byte, show that the ACT CRASE engine is capable of scoring essays in an operational administration.

## References

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

McGraw-Hill Education CTB. (2014). *Smarter Balanced Assessment Consortium field test: Automated scoring research studies in accordance with Smarter Balanced RFP 17*. Smarter Balanced Assessment Consortium. *https://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf*

## ABOUT ACT

ACT is a mission-driven, nonprofit organization dedicated to helping people achieve education and workplace success. Grounded in 60 years of research, ACT is a trusted leader in college and career readiness solutions. Each year, ACT serves millions of students, job seekers, schools, government agencies, and employers in the U.S. and around the world with learning resources, assessments, research, and credentials designed to help them succeed from elementary school through career.

For more information, visit act.org