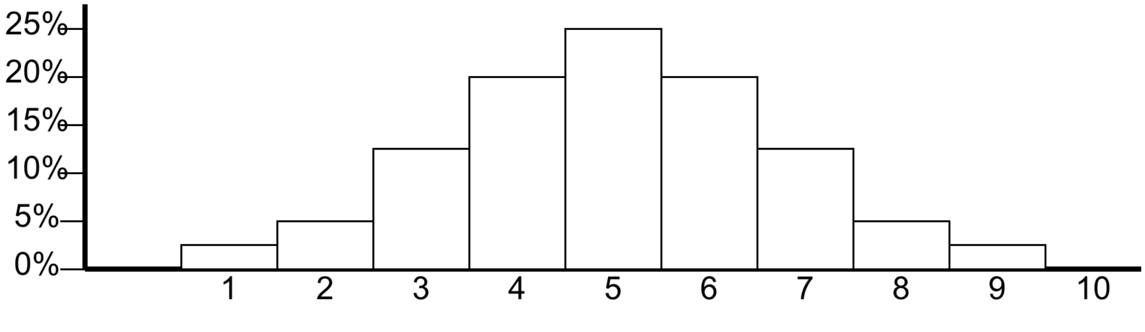
Science Fair Judges' Guide

Thank you for being a judge in the science fair. We would like to judge science fair projects accurately, consistently, and precisely so students can get the best possible feedback. We would like the judging process to be as intuitive as possible for you, so that you can fully express the quality of the science fair projects without becoming "bogged down" in recording too many numbers.

All of your ratings will be on scales from 0 (worst) to 10 (best). Exceptional (and rare) work of creativity and passion receives a ten (10). Zero (0) represents such poor work that we hope you will rarely need to use it. Three (3) represents minimal work of reasonably quality and seven (7) means solid work. You will judge each student on four (4) distinct dimensions using this exact same scale.

We came up with these 4 dimensions based on the feedback judges provided after previous years' science fairs. We asked what dimensions were most important to you and the result was a general consensus of 4 clusters. I continue to revise this rubric so if you would like to provide feedback after using this scoring system, please e-mail me (Dr. Kevin Grobman PerplexingQuestions.org).

More details about what 0, 3, 7, and 10 mean for each dimension are given below. Really use these benchmarks to decide where on the scale to place each student's work. Please do *not* adjust for other circumstances. Naturally, we expect that 6th graders to be less sophisticated than 12th graders. That means we understand when your average rating for 6th graders is lower than for 12th graders. When we combine your numbers for a composite rating, we will do scaling to make judging fair. Really using the benchmarks will be important when we look back at students over time. For example, we can follow how students progress from year to year. This will allow us to give schools feedback on, for example, how students are not mastering certain scientific skills as quickly as they learn others. Finally, really using the benchmarks is important because if judges consistently rate a student's work low or high, we will be able to give the student accurate feedback. Though you are free to use any number from 0 to 10 (e.g., 5.75), we recommend using whole numbers (e.g., 6) unless you feel that fine-grain differentiation between students is necessary.



Even though your primary goal should be giving accurate ratings on the scales, there are some patterns we expect (see histogram above). When your ratings differ, please consider if you should adjust your ratings or if this is a genuine instance where the ratings simply should be unusual. If you do not use at least 3 numbers in your ratings (e.g., you only give ratings of 6 and 7), then consider that you may not be noticing enough of the subtle differences between student's work. If you give more than 15% of students 8's, 9's & 10's, consider that you may be too lenient. If you give more than 15% of students 0's, 1's, & 2's consider that you may be too critical.

Within each dimension, you may feel something particularly important within your scientific field is not explicitly mentioned in the description of the benchmarks. Our rubric needs to be general because it has to apply equally well to, for example, physics and social science projects. The descriptions are meant only as summaries to help judges be consistent. They are *not* exhaustive descriptions to be taken literally. You can, and should, use your expertise about what (for example) "scientific rigor" means in your field and incorporate this into your rating.

Though your ratings use the benchmarks provided, you may be comparing projects in order to choose numbers within a small range (e.g., what distinguished a 5 and 6). Since you are an expert in your area, you are among those best equipped to make the subtle ratings. However, there are some situations that make comparisons difficult for even the best experts. Projects within a field are still on widely different topics. Adults help children to widely different extents. How can you "compare apples and oranges?"

> *Projects on Topics You Know Well:* If you happen to know more about some science fair project topics than others, you might inadvertently judge those projects more rigorously. When you know a topic well, you know more of the subtle incorrect things students. If students present something that is blatantly incorrect in a way you would detect in even topics you do not know especially well, the error should certainly lower the student's score. However, be mindful of if an error you see is subtle in a way you might not notice in other projects. Errors should always lower your rating, but do not deduct as much for subtle errors.
>
> *Different Amounts of Adult Help:* Sometimes children complete projects entirely on their own. Sometimes children get help from non-experts like their parents. And sometimes children work with experts, like professors. When rating projects, you need to look beyond adult's contributions and rate the child's work. When children figure things out on their own, their scores should rise for showing independence. When children find an adult who can help them with aspects beyond their ability, their scores should rise for showing skill at finding collaborators and learning something more advanced than their grade-level. But when adults do things for the child, the child should not get credit for the adult's effort. Here are some examples to help you differentiate "good" and "bad" help.
>
> *Bad:* An adult had something that interested him or her and told the child it would make a good project.
> *Good:* The child approached a professor studying something he or she likes and the professor helped refine the idea into a scientifically important hypothesis.
>
> *Bad:* An adult took measurements because they had the skill to be most precise (aside from using dangerous tools that children should never be expected to use).
> *Good:* A child wanted to measure something and he or she found an adult who recommended a precise tool. The adult and child measured together with the child doing as much as possible even if that meant less precise measurement.

Bad: An adult took the child's data and created polished results beyond the child's ability to understand.

Good: The child did analyses they knew about and sought adults with skills to produce better results. Even if the adult did the analysis, the child understands what they did and why.

The language for describing the benchmarks may seem unusual to you because we needed to have consistent terms for judges in every scientific field. The following is a list of terms we feel could be most confusing.

Testable Hypothesis: A statement that may be true or false and a statement where evidence that can help us decide. For example, "The sky is purple" is a hypothesis and looking up shows it is false. "All single men are bachelors." is not a hypothesis because it's just true by definition. If a student has a hypothesis and it could not possibly be wrong, no matter what their study showed, then either their study was irrelevant to the hypothesis or the hypothesis was not testable (not scientific). (Note: for judging mathematics, "looking for evidence" may need to be changed into "writing a proof").

Operational Definitions & Theoretical Constructs: A theoretical construct is something we cannot directly know but something we can approximate by measurement (our operational definition). For example, gravity is a theoretical construct and measuring the acceleration of objects we drop is a way of operationally defining it. Intelligence is a theoretical construct and IQ tests are a way we operationally define it.

Confound: When relating two measures, other theoretical constructs might be important to interpreting the results. Confounds are the unmeasured third variables of particular importance or variation between conditions that was not part of an experimental manipulation. Confounds bias interpretation of results. For example, suppose a student looked for gender differences in memory and found girls have better memory than boys. Her measure of memory was seeing photographs of students in distinct clothing and then asking participants to match new photographs of the students with new photographs of their clothes. Maybe the difference in performance is because of memory or maybe it is because of the confound of interest in clothing (a stereotypical interest of girls).

A computerized spreadsheet will determine the composite rating so you do not need to write anything beside 4 numbers. Each judges' ratings for each category and grade level (e.g., senior level physics) will be considered separately for awards. Statistical methods (e.g., z-scores) may be incorporated into the final ranking of science fair projects to adjust for individual differences between science fair judges.

This guide focuses on judging accurately and students appreciate being asked challenging questions as long as you're also friendly. In fact, how students perceive you as fair, rigorous, friendly, and warm is the best predictor of students' interest in doing a project again next year according to our research (Dr. Kevin Grobman). You play the most crucial role in children's experiences at the science fair. Thank you for being part of it.