Automatic Evaluation

How would you design a system to automatically evaluate computing problems for a competition like the BIO? You are free to mention anything you feel would be relevant, such as the types of tests you would perform, how you would deal with different programming languages, the varieties of tasks you could (or could not) evaluate, submissions to the system and their feedback, and how the system is affected by multiple competitors. You should mention the feasibility and limitations of your suggestions, but there is no need to go into overly specific implementation details.