Science Fair “Fairness”

Douglas Leroy Adams

University of Utah

Department of Educational Psychology

Master of Statistics Final Project

Contents

Introduction

Science fairs are an integral part of our culture here in the United States. It seems almost everyone has experience with them in some form or another. Many have created their own experiments and have displayed them in a science fair held by their public school. Others have attended science fairs to support others such as their children or friends. Parents, teachers, and other mentors become involved as they help students with projects. Still, others have been recruited to fulfill the role of a judge. In any case, virtually everyone can relate to science fairs. They're part of our cultural experience. One might even venture to say that they share equal status with other traditions that decorate the lives of our youth, much like dodge ball in gym class, movies with popcorn, and senior prom.

The Beginnings of Science Fairs

Some might say America saw its most rudimentary instances of scientific exhibition as early as the 1800s, when county fairs displayed the latest farming inventions (Cox, 2007). But when did the science fair as we know it begin? How did it become an academic competition, and how did it become so popular?

The men behind science fairs. Edward Willis Scripps was a journalist and businessman who lived in the late 1800s and early 1900s. His career in journalism actually started out when he began working at his brother's newspaper as a young man. As the years passed, Scripps became more interested in managing newspaper companies rather than simply working in them and, with financial help from his family, he began to acquire or create from scratch a number of local newspapers. Although he would own these smaller companies, Scripps would allow them a great degree of license in how they ran themselves. These businesses eventually grew into a large empire which is known today as the E. W. Scripps

Company ("E. W. Scripps," 2009).

Later in life, Scripps would meet William Emerson Ritter. Ritter was a scientist whose specialty was in biology and zoology. He took up several teaching positions throughout his academic career. During this period he had the opportunity to read and study science in great detail. Ritter took advantage of these opportunities. By 1881, he had obtained a teaching position that afforded him the time to study the sciences with more intensity than he had before. He became enthralled with the way science could view the world around him, and he began to feel that an extension of this viewpoint to worldwide issues would be the solution for much of the problems and suffering in the world. Thus William Ritter became a vigorous advocate of scientific thought and its universal application (“William Emerson Ritter,” 2009).

In 1903, E. W. Scripps and another family member donated money to help establish the Marine Biological Association of San Diego, which still exists today as the Scripps Institution of Oceanography. While at first Scripps wasn't excited about the endeavor, he later became friends with its founder, William Emerson Ritter (“E. W. Scripps,” 2009). Together they began sharing ideas about projects ongoing at the association. This sparked within E. W. Scripps a deeper interest in science – and now, because both men had a strong friendship and a common interest, it would only be a matter of time before Scripps' financial resources and Ritter's scientific expertise would accomplish even greater things.

The Science Service and Science Talent Search. And so they did. In 1921, they formed the Science Service, an organization with an objective to bring advances in science to the public eye. As some in the scientific community did not believe this would be very constructive, the view of Scripps and Ritter was in the minority. However, they felt strongly that everyone should have the chance to be informed of current developments in science

("E. W. Scripps," 2009). They began publishing a newspaper called the Science News – a newspaper for the layperson, and printed by the Science Service itself. This opened the door for the public to stay apprised of the current scientific advancements of the day.

But what transpired a couple of decades later is actually why we might well consider these gentlemen the founders of modern science fairs as they exist today. Although Scripps passed away in 1926, Ritter maintained his involvement in the Science Service, the passion the two men held for scientific propagation remained intact, and the Science Service continued onward. In 1941, the organization worked with the American Institute of the City of New York to initiate local "science clubs" across the nation. These clubs were created to encourage students to participate in the sciences more, and with the hope that eventually more of these students would aim their careers in that direction (History of Society for Science & the Public, 2009). Even more significantly, however, the Science Service would collaborate with Westinghouse in 1942 to hold the first ever "Science Talent Search." This was the event that began science fairs. The Science Talent Search operated much the same way a modern science fair does; all of the main elements of a science fair existed in the first Science Talent Search: students across the nation were invited to submit projects for a competition, and the winners were awarded various prizes. Even today the Science Talent Search is held each year – now with Intel – and prizes from laptops to scholarships are awarded.

Gaining popularity. The 1950s was an eventful era and witnessed a great increase in the popularity of science fairs. Advances in nuclear physics, space exploration, computers, and television were exciting, and the general public became more interested in science and technology. In the 1960s and 1970s, science fairs started to become much more commonplace in schools (R. and E. Adams, personal communication, May, 2009).

Now we see science fairs held regularly across the United States and the rest of the world. Parents and teachers enjoy helping students while students have fun, learn, and even grow in how they think about the world (Smith, Maclin, Houghton & Hennessey, 2000).

Judging Science Fairs

Many if not most science fairs host multiple categories, sometimes being based on student demographics such as age or school division. Other times categories are created to divide student projects into subject areas, ensuring that a broad sample of scientific disciplines are represented. Categories may also be created based on different attributes of a project such as originality, personal effort, level of complexity, or clarity in articulation. This can be done to allow for more students to be recognized and awarded for their work. But however the event is organized, there is one element common among all science fairs: judging.

Judges are recruited, usually from local communities, to score student projects based on some measure of quality, inventiveness or creativity. Of course, definitions of quality or creativity can fluctuate, and can surely be debated on many different levels. So with a little contemplation one might well wonder: How are science fairs judged fairly? Or perhaps the first question is actually: Are they even judged fairly in the first place?

"Fairness." That depends on what 'fair' means. Is one aspect of fairness more important or more measurable than another? To statistically probe a science fair for bias, which will be the purpose of this study, it will be practical to consider variables that are easily scaled or that are already measured, such as the student's age group, the judge's area of expertise, or the scores awarded to a project. Aside from basic student information, then, most of the statistics that can be derived from a science fair are about the scores and judges themselves. It may also be intuitive to note that the fairness of any judgment will have to do with the judge – so any analysis of fairness will focus on various aspects of judging and

judges themselves.

How does a community prepare a judge to judge fairly? There are those who feel that a standard should be followed for how judges interact with students: that is, that there should be a dictation on some minimum time allotted for a judge or group of judges to spend with each student. The Greater San Diego Science and Engineering Fair gives each student fifteen to twenty minutes with a team of five or six judges – which teams have been carefully 'balanced' with members selected from a healthy variety of scientific disciplines and experience levels (Frederickson & Mikkelson, 1979). This could be advantageous. From Classical Test Theory, we know that a larger group of judges will tend to produce an average score that will be more accurate in its measure of the science project's “true score,” having less random error than would an average score produced by a smaller group of judges (Crocker & Algina, 1986).

Another tactic many communities utilize is to lay out specific criteria for its judges to follow when rating a student project. For example, the Salt Lake Valley Science and Engineering Fair holds a training session for its volunteer judges and gives them a rubric from which to make evaluations (J. Ostrander, personal communication, January, 2009). This is an attempt to ensure that the criteria used in appraising each science project are the same for every judge.

But even after such pains are taken to ensure that science fairs are judged with high standards of quality and equality, how do we know those judges are scoring fairly?

Judges. Of course the premise for any concern regarding whether science fairs are scored fairly stems from the inescapable fact that the mechanism for scoring science fair projects is human judgment. With all the efforts to standardize judgment and use an enlightened set of

criteria for assessment, there are elements of human judgment involved in that assessment. However well the fair is administered and however well the judges have been trained, there will be errors in judgment. What happens when some judges are in a better mood than other judges on the day of the fair? Or what if a judge is assigned to a student project that shares topics in common with that judge's field of interest? Would she give a generous score to the student because the project is very interesting to her or, if she were aware that such an effect were entirely possible, would she try to score more strictly in order to “balance out” that effect? Or would a strict judgment reflect the fact that her specialized knowledge has given her the insight to see flaws that judges with other specialties would have missed? Or, could it be that scores she awards to projects in her area of expertise actually vary more than those she gives outside it? Maybe her intimate knowledge of a topic allows her to judge more discriminately in it, shifting the score further up or down according to more subtle details other judges would not see.

Another angle to consider is the fact that judges do not come to the table with the same set of academic experiences, or even level of academic achievement. Will the ones with higher educational backgrounds tend to be more harsh because they are conditioned to expect more, or will they realize this and overcompensate by not expecting as much as other judges? Or would the scores they give vary more or less than those given by judges with less advanced degrees?

Similar to questions such as these is one involving a possible age effect. Are younger students more likable in the eyes of some judges than they are to other judges? And again, what if a judge realizes that younger children are more adorable to her, and she feels obligated to 'correct' her judgment for emotional bias? Will she under or over-correct her score while taking into account her emotions?

While examining specific questions such as these, it may also be informative to investigate the attributes of judges who score typically lower or higher than those of other judges. Do judges who give lower scores than other judges on average have certain characteristics in common with each other? Do judges who score higher on average tend to be those of a certain profession, or do they work for smaller or larger companies?

These are the questions this study will address. For convenience they are listed more parsimoniously below:

For judges who judge both inside and outside their own areas of expertise:

- Is there a significant difference between the scores they give inside versus outside their areas of expertise?
- Do scores vary significantly more or less when given inside or outside areas of expertise?

When judges are grouped by level of education:

- Is there a significant difference in scoring between groups?
- Do all the scores given by a single judge vary significantly more or less depending on the group to which the judge belongs?

For younger versus older students:

- Is there a significant difference between scores?

Looking at deviations of judges' scores from projects' mean scores:

- Are there common characteristics among judges who score higher or lower than the average scores given by other judges of the same projects?

While we don't have the resources to answer these questions for all science fairs, of course, we can turn to the local arena to gain some insights.

The SLVSEF

The Salt Lake Valley Science and Engineering Fair (“SLVSEF”), mentioned previously, began in 2002 and is held annually in Salt Lake City, Utah by the Utah Science Center and the University of Utah. It attracts participating students from schools all over the Salt Lake Valley. Elementary, junior and senior high schools are all invited. Judges come from a wide variety of educational backgrounds, professions and scientific interests. The organizers of the SLVSEF have provided four years of quantitative data from the fair (2006, 2007, 2008 and 2009) containing information about students, judges and project scores. These data will serve as the data-set for testing the previously stated research questions.

Methods

Data Description and Preparation

In the SLVSEF, each judge scores multiple student projects, and each student project is scored by multiple judges, although not all possible pairings of judges and student projects occur. Judges initially complete a form which allows them to designate their first, second and third favorite subjects in which they would prefer to judge. Officials attempt to assign judges to projects belonging to these subjects, but most judges are also assigned to other projects as well.

Here is an overview of the four years of original data:

2006	2007	2008	2009
46 schools	59 schools	62 schools	61 schools
82 teachers	94 teachers	110 teachers	162 teachers
191 student projects	258 student projects	217 student projects	291 student projects
175 judges	162 judges	124 judges	178 judges

The data received was in spreadsheet format, consisting of 3 worksheets for each year: student projects, judges, and a score table matching every score with a judge ID and a project ID. For every student project, the information recorded consists of the individual student identifiers (name and science fair ID), the name of the project, the category in which it was placed (Behavioral Science, Chemistry, etc.), the teacher or mentor involved in the project, and the student's school and division (elementary, junior or senior). The available data about each judge consists of individual identifiers (name and judge ID),

degree level, areas of expertise, years of experience in teaching or industry, any previous judging experiences, and the judges' first, second and third choices of categories in which they would prefer to judge. For 2008 and 2009, the judges' place of residence and current employment is also recorded.

Because some of the students' mentor names were written differently for the same person (for example, "Mr. Ken Baker" and "Kenneth Baker" of the same school), all such entries in my copy of the data (the "processed data") were edited so the mentors' names would be consistent. At times the school's website was consulted when it was not obvious whether there were in fact two mentors with similar names. This was also the case with school names. This was done to avoid analysis errors that would have otherwise occurred if there were more mentors or schools in the data than were actually involved in the fair.

The 2006 judge data was recorded using fourteen scientific categories for which judges could claim an expertise or propensity. These were:

behavioral science
biochemistry
botany
chemistry
computer science
earth science
engineering
environmental management
mathematics
medicine
microbiology
physics
space science
zoology

The original 2007 data did not use these categories. The 2008 judge spreadsheet listed them but did not use them, while the 2009 judge spreadsheet did. Since those categories seemed logical and were the most commonly used categorization scheme throughout all four years of the original data, it was decided that the best way of classifying judges would be to use this categorization scheme for all four years of data. Where judges had not been

categorized using those fourteen categories in the raw data, the processed data had to be adjusted so that all judges would be classified into those categories. This was accomplished by examining each of those judges by their education, interests and employment, and then assigning the judges to as many of those categories as seemed appropriate. For example, one 2008 judge lists his expertise areas as “Meteorology / Hydrology,” so he was assigned into the 'earth science' and 'environmental management' categories. Another listed her expertise areas as “biochemistry, structural biology and chemistry” and was assigned into the 'biochemistry,' 'chemistry' and 'microbiology' categories.

The original data included some judges whose educational or professional degrees were pending completion. For the sake of simplicity, the processed data lists those judges as if they had completed those degrees. No distinction is made in the processed data between a Bachelor of Arts and a Bachelor of Science, or between a Master of Arts and a Master of Science. Those judges who were originally listed as medical doctors, pharmacists, or lawyers were listed in the processed data with those who have PhDs.

Descriptions of Analyses

Judges who judge both inside and outside their own areas of expertise. Some of the SLVSEF judges were only assigned student projects that matched their own area of expertise, while other judges were only assigned projects outside their area of expertise. The rest of the judges had the opportunity to judge student projects both in and out of their area. Considering this, a decision was made to use only the scores of the judges who judged both in and out of their areas of expertise. Were we to attempt to include all judges in this analysis, we would need to consider 'expertise' as a factor with two levels: judging 'within' or 'without' area of expertise. But because so many judges did not have the opportunity to judge both in and out of their area, while so many others did, neither a “within subjects” nor a “between subjects” approach would be appropriate. However, using only those judges who judged both in and out of their area, we can meaningfully compare scores resulting from both types of judging (in and out of the judge's area of expertise). By averaging scores for both types of judging for each judge, we will have two mean scores for every judge. These will of course represent an overall bearing on how that judge scored in and out of her expertise. Essentially, this is done to avoid nesting effects from other factors such as 'student project'; in other words, by averaging scores we will not be dealing with how student projects within each judge and judgment type may or may not be correlated.

Using the same reasoning for grouping scores together as a mean within each judge and type of judging, standard deviations will also be utilized. These will allow us to compare whether the dispersion of scores, in versus out of judges' areas of expertise, is significantly different.

This data can be examined using a correlated samples t-test, where the subjects are the

judges, and the two factor levels will be judging 'within' and judging 'without' area of expertise. In the case where the data does not meet the assumptions, or standards, required for the use of a t-test – namely, normality and homogeneity of variance – the Wilcoxon Signed Rank (Matched Pairs) test will be used as a nonparametric alternative.

Judges grouped by education level. The education level of each judge is reported as either a bachelor's degree, master's degree, or doctorate degree. Although an ideal data set might have all judges judging student projects from all divisions (elementary, junior, and senior), most judges in the existing data set judge exactly two out of the three student divisions: most of the time either elementary and junior, or junior and senior. Very few judges actually judge projects from all three divisions. Because of this, an ideal study cannot be performed in which scores are nested within student divisions, and where student divisions are nested within judges.

For this question then, an approach similar to the one made in the previous question will be helpful. Again, it will make sense to look at the average scores given by the judges. By ignoring student division (elementary, junior or senior) and looking only at the judge's education level as the factor under consideration, we will at least be able to tell whether there is some effect on scores related to education level.

As with the question about areas of expertise, this analysis will be carried out twice for each year of data: once using the means and once using the standard deviations of each judge's scores. Again, this will allow us to examine whether or not the judges' education level affects the mean of the scores they give, and also whether or not their education level affects the amount of dispersion in the scores they give.

An ANOVA (Analysis of Variance) will be used to examine this question. The analysis will

utilize a "between subjects" design, as judges are naturally grouped according to their education level. Where the data do not meet the assumptions of an ANOVA, the nonparametric Kruskal-Wallis test will be used. All available judges from each year will be used.

Students of different ages. This question may seem straightforward at first: whether younger students tend to get higher scores, for instance, appears to be a simple tally. Yet that tally, though easy to perceive and calculate, may not give an answer as enlightening as an analysis that takes into account groupings within the data. For example, what if all students consistently received high scores in one particular school while other students in the same age group but in a different school received lower scores on average? The one school of high scoring students may throw off the tally so as to misrepresent the score level of the majority of students in that age group.

An analysis carried out as such is termed a "disaggregated analysis" (Twisk, 2006) because it does not take into account the possible clustering, or aggregation, of the scores of students within schools. In other words, it yields a "disaggregation bias" in the results.

On the other hand, consider that the effect of schools is taken into account, but this time no individual student scores are used – only an average of all scores from the same school. In this case all information regarding the variation of scores within each school is lost, where a comparison of these might have been helpful. This analysis is called "aggregated analysis" because the data is grouped, or aggregated, and may produce an "aggregation bias" (Twisk, 2006) by not considering patterns of the data within groups (student scores within schools).

For this question, then, hierarchical linear modeling (HLM) will be useful because it will take into account the grouping of students within schools. The independent variables of this

hierarchical model will be the school each student attends, and the student's division (elementary, junior high, or senior high), which will determine the student's age group. The dependent variable is the student's average score. The model will thus consist of two random levels: school and students' average scores. The division in which the student is competing is a fixed factor at the student level.

The linear model used will be a "random intercepts" model. This means that a separate regression line is essentially calculated for every school, each with its own intercept, and that the regression slopes are the same for every school. While having parallel regression lines for every school assumes that Division affects the scores within all schools equally, this allows us to test for Division effects much more easily than would a model with both random intercepts *and* random slopes.

The model used for this question will be:

$$y_{ij} = \beta_0 + \beta_1 I(\text{Junior}) + \beta_2 I(\text{Senior}) + u_j + e_{ij}$$

where

$\mathbf{y_{ij}}$ = dependent variable (student's average score)

$\mathbf{\beta_0}$ = predicted mean for all scores at 'Elementary' division level

$\mathbf{\beta_1}$ = predicted mean offset (from Elementary mean) for all scores at 'Junior' division level

$\mathbf{\beta_2}$ = predicted mean offset (from Elementary mean) for all scores at 'Senior' division level

I(Junior) = indicator function: 1 if division is 'Junior' and 0 otherwise

I(Senior) = indicator function: 1 if division is 'Senior' and 0 otherwise

$\mathbf{u_j}$ = contribution of school j

$\mathbf{e_{ij}}$ = contribution of student i within school j,

and where $\mathbf{u_j} \sim N(0,\text{variance between schools})$ and $\mathbf{e_{ij}} \sim N(0,\text{variance within schools})$.

Because Division is the only information available about student age, the question of whether the scores of younger versus older students differ will be addressed by testing to see if scores differ by level of Division. This will involve comparing the values of the likelihood functions of two linear models, the simpler of which is said to be "nested" within the more complicated (Pinheiro & Bates, 2002). For this case the more complicated model will be one in which Division is modeled, and nested within that will be the simpler model where Division is not taken into account. The values of the likelihood functions of these two models, given the data, will be compared using a likelihood ratio test, which follows a chi-square distribution with the same number of degrees of freedom as the number of additional parameters being estimated in the more complicated model. If the result of the likelihood ratio test implies that the data is better fit using the more complicated model, we can infer that the influence of Division effects on scores is not negligible.

The nested model used for the likelihood ratio test will be:

$$y_{ij} = \gamma_{00} + u_j + e_{ij}$$

where

$\mathbf{y_{ij}}$ = dependent variable (student's average score)

$\boldsymbol{\gamma_{00}}$ = grand mean of all scores

$\mathbf{u_j}$ = contribution of school j

$\mathbf{e_{ij}}$ = contribution of student i within school j,

and again where $\mathbf{u_j} \sim N(0,\text{variance between schools})$ and $\mathbf{e_{ij}} \sim N(0,\text{variance within schools})$.

These parameters will be estimated using Restricted Maximum Likelihood procedures. The model will then be refitted using full Maximum Likelihood procedures in preparation for the likelihood ratio test. Although full Maximum Likelihood computations are required for the likelihood ratio test, Restricted Maximum Likelihood estimates will be reported, as full

Maximum Likelihood calculations tend to underestimate parameters much of the time (Pinheiro and Bates, 2002).

Because hierarchical linear modeling is like a more generalized and flexible implementation of Ordinary Least Squares (OLS) regression, it allows questions to be asked of the data at different levels (http://www.cmm.bris.ac.uk/lemma/). In addition to reporting on whether Division is significant for each year of science fair data, this paper will also seek to answer whether or not School as a random effect is significant; in other words, whether the grouping of students within schools accounts for a significant amount of the variance in the model. This will also be accomplished using a likelihood ratio test, but interestingly enough one that does not require that the models be fit using full Maximum Likelihood; the models fit using Restricted Maximum Likelihood will be used. The test will involve comparing the likelihood values of the original model versus one that does not include the random effect $\mathbf{u_j}$.

Deviations of individual judges. The author will attempt to find attributes common to judges who score higher on average, and those who score lower on average, than the mean scores of the projects they judge. First, the average score for every project will be calculated. This will enable a comparison between every judge's score for each project he or she judges against the average score awarded to that project. Once those deviations are recorded, the mean deviation of each judge from the average score given to each project judged is calculated.

The analysis will take an interesting turn at this point: no statistical test is really associated with the finding of patterns such as these, yet this loose kind of exploration of the data allows the author a certain freedom to gauge perhaps a larger picture of judging trends than even a series of statistical tests might.

Software. For the analyses, all four years of data will be exported from SQLite databases and imported into an R workspace. Data will be aggregated into mean scores and standard deviations of scores, and descriptive statistics about these data will be obtained by again aggregating the aggregated data in a similar manner. The functions used for t-tests, Wilcoxon Signed Rank tests, ANOVAs and Kruskal-Wallis tests are included with the base installation of R. The "lme4" package, by Douglas Bates and Martin Maechler (2010), will be used for hierarchical linear modeling. The "languageR" package (Baayen, 2009) will be utilized for further examination of confidence intervals, and the "RLRsim" package by Fabian Scheipl (2010) will be used to test random effects.

Correlation across years. A large number of students, teachers and parents overlap in their involvement in the science fair from year to year. Because of this, answering a question using a single analysis across all four years of data combined would be inappropriate. Each of the three questions will be addressed four times – once for every year of data.

Data limitations. A more ideal situation of course would be the case in which enough data exists to analyze all three questions using hierarchical linear modeling. The question about judges scoring in their own area of expertise would be a prime target for hierarchical modeling if all levels of data for every judge were to exist, and if enough students were scored by every judge. Demographic information about students such as teacher and school could be considered in a hierarchical model because there would be enough students scored by each judge. Similarly for the question about the education level of the judges: if enough students were scored by each judge, perhaps demographic information could be modeled using a hierarchical linear model.

Results

Area of Expertise Data

Descriptive statistics. The following table gives basic descriptive statistics for the Area of Expertise data where both sets of scores (in and out of areas of expertise) for every judge were averaged.

Year	N	Means of Means	StDevs of Means
2006	in : 51 out : 51	in : 78.58822 out : 79.50857	in : 7.531225 out : 8.506835
2007	in : 45 out : 45	in : 77.43719 out : 78.59902	in : 8.612623 out : 7.664810
2008	in : 52 out : 52	in : 78.11692 out : 79.13322	in : 9.826044 out : 9.780263
2009	in : 53 out : 53	in : 77.65586 out : 79.05400	in : 7.626987 out : 7.161549

In the table below, the standard deviation of all scores in a 'set' is considered. In other words, the descriptive statistics are for standard deviations instead of score means. Notice that the standard deviations in the table above are different from the mean standard deviations in the table below. The standard deviations reported in the table above are calculated across the scores of all judges for each group, whereas the means below are calculated using the standard deviations from individual judges. (The standard deviation of all scores is not the same as the average of standard deviations from smaller subgroups of those scores.)

Year	N	Means of StDevs	StDevs of StDevs
2006	in : 51 out : 51	in : 10.957730 out : 9.524107	in : 5.803114 out : 5.176137
2007	in : 45 out : 45	in : 11.57460 out : 11.35842	in : 5.088999 out : 5.973106
2008	in : 52 out : 52	in : 10.41314 out : 10.19923	in : 4.806166 out : 5.372600
2009	in : 53 out : 53	in : 11.39699 out : 10.53893	in : 4.484539 out : 5.706041

Determining the appropriate statistical tests. The Area of Expertise data was tested to see which years passed the assumption of normality using the Shapiro-Wilk test, and the homogeneity of variance assumption using the Fligner-Killeen test. This was done for the data where the means of every judge and area-of-expertise combination are considered as data points, and also for the data where the standard deviations are considered as data points. The null hypothesis for the Shapiro-Wilk test is that the data is Normal, and the null hypothesis for the Fligner-Killeen test is that the variances are homogeneous. The test values, degrees of freedom (denoted "dof") and p-values for these tests are shown in the tables below for every year, and the results of the Shapiro-Wilk tests are shown for both types of judging (in and out of areas of expertise).

Pre-Tests on Data using Means as Data Points							
			Shapiro-Wilk Test		Fligner-Killeen Test		
Year	N		Test	p-value	Test	dof	p-value
2006	in : out :	51 51	W=.9577 W=.989	.06689 .9157	X^2=.5329	1	.4654
2007	in : out :	45 45	W=.9722 W=.9604	.3471 .1262	X^2=1.2879	1	.2564
2008	in : out :	52 52	W=.9226 W=.9522	.002347 .03615	X^2=.0783	1	.7797
2009	in : out :	53 53	W=.9531 W=.9635	.03666 .1048	X^2=.218	1	.6406

Pre-Tests on Data using Standard Deviations as Data Points							
			Shapiro-Wilk Test		Fligner-Killeen Test		
Year	N		Test	p-value	Test	dof	p-value
2006	in : out :	51 51	W=.9453 W=.9288	.02013 .004496	X^2=.0053	1	.9417
2007	in : out :	45 45	W=.957 W=.9541	.09353 .07284	X^2=.911	1	.3398
2008	in : out :	52 52	W=.9639 W=.9299	.1160 .004461	X^2=.0588	1	.8085
2009	in : out :	53 53	W=.9596 W=.9187	.07057 .001497	X^2=1.742	1	.1869

Using alpha = .05, the only data that appears to be fit for t-tests would be the 2006 means, 2007 means, and 2007 standard deviations.

Test results. Using the null hypothesis that there is no difference between judges' mean scores when judging in versus judging out of their areas of expertise, test values, degrees of freedom and p-values for the previously determined appropriate tests are shown below.

Tests on Data using Means as Data Points				
Year	Test Used	Test	dof	p-value
2006	paired t-test	t=.8119	50	.4207
2007	paired t-test	t=.8734	44	.3872
2008	Wilcoxon Signed Rank	V=756	NA	.3859
2009	Wilcoxon Signed Rank	V=852.5	NA	.2269

Assuming t-tests are considerably robust when it comes to violations of normality, t-tests for 2008 and 2009 are shown as well.

Tests on Data using Means as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2008	paired t-test	t=.9373	51	.353
2009	paired t-test	t=1.4721	52	.1470

Using the null hypothesis this time that there is no difference between the standard deviations of judges' scores when judging in versus judging out of their areas of expertise, results from the appropriate tests are shown below.

Tests on Data using Standard Deviations as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2006	Wilcoxon Signed Rank	V=527	NA	.2040
2007	paired t-test	t=-.1977	44	.8442
2008	Wilcoxon Signed Rank	V=630	NA	.5942
2009	Wilcoxon Signed Rank	V=538	NA	.1171

Again, t-tests are also carried out on the data subsets where Wilcoxon Signed Rank tests were initially used.

Tests on Data using Standard Deviations as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2006	paired t-test	t=-1.444	50	.1549
2008	paired t-test	t=-.2696	51	.7885
2009	paired t-test	t=-1.185	52	.2416

With alpha = .05, there is no conclusive evidence in the data that judges judge differently – either on average or in the variation of scores they give – when they judge in or out of their

own areas of expertise. This is consistent for all parametric and non-parametric tests. There is, however, possibility of cumulative Type I error due to the fact that more than one statistical test has been performed on the same data.

Education Level Data

Descriptive statistics. The following table gives descriptive statistics for the Education level data, where all the scores for every judge were averaged.

Year	Total N	N	Means of Means	StDevs of Means
2006	111	55 23 33	Bachelor's : 79.4205 Master's : 77.6337 Ph.D. : 80.0114	Bachelor's : 7.3563 Master's : 6.1930 Ph.D. : 6.6694
2007	124	56 26 42	Bachelor's : 78.6601 Master's : 79.6074 Ph.D. : 77.5576	Bachelor's : 6.8352 Master's : 6.8593 Ph.D. : 8.0009
2008	119	43 30 46	Bachelor's : 78.5923 Master's : 78.1912 Ph.D. : 76.8794	Bachelor's : 7.2250 Master's : 8.7384 Ph.D. : 8.2208
2009	142	65 34 43	Bachelor's : 79.7130 Master's : 77.6257 Ph.D. : 79.3693	Bachelor's : 7.7768 Master's : 6.9616 Ph.D. : 9.9032

The table below, similar to that of the second Area of Expertise descriptive table, describes the Education level data where the standard deviation of a judge's scores is considered the data point for that judge. As with the Area of Expertise data, note that the standard deviations in the table above are different from the mean standard deviations in the table below.

Year	Total N	N	Means of StDevs		StDevs of StDevs	
2006	111	55	Bachelor's :	10.3148	Bachelor's :	4.4198
		23	Master's :	11.8373	Master's :	3.7223
		33	Ph.D. :	11.3874	Ph.D. :	4.6893
2007	124	56	Bachelor's :	10.9424	Bachelor's :	3.9599
		26	Master's :	11.7155	Master's :	3.6169
		42	Ph.D. :	11.8112	Ph.D. :	5.3516
2008	119	43	Bachelor's :	10.9671	Bachelor's :	3.9897
		30	Master's :	11.6891	Master's :	3.9939
		46	Ph.D. :	11.0682	Ph.D. :	4.7969
2009	142	65	Bachelor's :	10.8768	Bachelor's :	4.7960
		34	Master's :	11.4978	Master's :	4.0236
		43	Ph.D. :	9.9243	Ph.D. :	4.3643

Determining the appropriate statistical tests. The Education Level data was also tested to see which years pass the normality and homogeneity of variance assumptions, again using the Shapiro-Wilk and Fligner-Killeen tests, respectively. And as before, these pre-tests were performed on both instances of the processed data: once where each judge's scores were aggregated into means, and once more where each judge's scores were aggregated to the standard deviation.

Pre-Tests on Data using Means as Data Points							
			Shapiro-Wilk Test		Fligner-Killeen Test		
Year	Total N	N	Test	p-value	Test	dof	p-value
2006	111	55	Bachelor's : W=.9724	.2364	X^2=1.269	2	.5302
		23	Master's : W=.9574	.4121			
		33	Ph.D. : W=.9743	.6065			
2007	124	56	Bachelor's : W=.9837	.6486	X^2=.9159	2	.6326
		26	Master's : W=.9689	.596			
		42	Ph.D. : W=.9756	.4974			
2008	119	43	Bachelor's : W=.9844	.8172	X^2=1.2603	2	.5325
		30	Master's : W=.9259	.0383			
		46	Ph.D. : W=.9822	.6974			
2009	142	65	Bachelor's : W=.9776	.2847	X^2=4.9569	2	.0839
		34	Master's : W=.9674	.3947			
		43	Ph.D. : W=.9572	.1098			

Pre-Tests on Data using Standard Deviations as Data Points								
			Shapiro-Wilk Test			Fligner-Killeen Test		
Year	Total N	N		Test	p-value	Test	dof	p-value
2006	111	55 23 33	Bachelor's : Master's : Ph.D. :	W=.9182 W=.9357 W=.9379	.0011 .1454 .0589	X^2=1.1554	2	.5612
2007	124	56 26 42	Bachelor's : Master's : Ph.D. :	W=.9757 W=.9780 W=.9496	.3161 .8299 .0623	X^2=1.3773	2	.5023
2008	119	43 30 46	Bachelor's : Master's : Ph.D. :	W=.9621 W=.9806 W=.9596	.1654 .8416 .1104	X^2=.5428	2	.7623
2009	142	65 34 43	Bachelor's : Master's : Ph.D. :	W=.9148 W=.9747 W=.9588	.0003 .6005 .1253	X^2=.1029	2	.9498

Using alpha = .05, the only data for which Analysis of Variance would ***not*** be appropriate would be the 2006 standard deviations, 2008 means, and 2009 standard deviations.

It is interesting to note how most of the data, even the standard deviations, is normally distributed. One might understandably expect the means to be normally distributed because the Central Limit Theorem may lead to such an intuition – even though these means are not sampled from the same theoretical distribution. But that most of the standard deviations are normally distributed is a pleasant surprise from an analytical perspective.

Test results. Examining the Education Level data with the null hypothesis that a judge's education level does not affect the mean score given by that judge, results for the previously determined appropriate tests are given below.

Tests on Data using Means as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2006	between-subjects ANOVA	F=.8403	2	.4344
2007	between-subjects ANOVA	F=.6704	2	.5134
2008	Kruskal-Wallis	X^2=1.377	2	.5023
2009	between-subjects ANOVA	F=.7352	2	.4813

As before, the parametric test is also carried out on the subset of data on which the non-parametric test was initially used, using the same level of significance (alpha = .05).

Tests on Data using Means as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2008	between-subjects ANOVA	F=.5488	2	.5792

Now, examining the Education Level data with the null hypothesis that a judge's education level does not affect the dispersion of scores given by that judge, test values, degrees of freedom and p-values are shown below for the appropriate tests.

Tests on Data using Standard Deviations as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2006	Kruskal-Wallis	X^2=3.978	2	.1368
2007	between-subjects ANOVA	F=.5482	2	.5794
2008	between-subjects ANOVA	F=.2755	2	.7597
2009	Kruskal-Wallis	X^2=3.0296	2	.2199

The parametric tests are again carried out on the subsets of data that did not strictly pass the tests of normality at alpha=.05.

Tests on Data using Standard Deviations as Data Points				
Year	**Test Used**	**Test**	**dof**	**p-value**
2006	between-subjects ANOVA	F=1.22	2	.2993
2009	between-subjects ANOVA	F=1.222	2	.2977

At alpha = .05, the data contains no conclusive evidence that a judge's education level has any effect on the mean score or dispersion of scores given by that judge. The parametric and non-parametric tests agree, although the author must again point out the possibility of accumulating Type I error.

Age Data

Descriptive statistics. It may be meaningful to give descriptive statistics for schools, as students are grouped within schools at level 2 of the model. Because of the number of schools involved in each year of the fair, these statistics are given in the appendix.

Parameter estimates. For reference, the linear mixed model used for this question is repeated below.

$$y_{ij} = \beta_0 + \beta_1 I(\text{Junior}) + \beta_2 I(\text{Senior}) + u_j + e_{ij}$$

Estimates for these parameters, calculated using Restricted Maximum Likelihood procedures, are as follows.

		2006 Estimate	2007 Estimate	2008 Estimate	2009 Estimate
Total # of students		183	233	217	260
Elementary mean	β_0	79.24	81.18	80.40	80.31
Junior offset from Elementary mean	β_1	-1.47	-4.61	-6.19	-3.25
Senior offset from Elementary mean	β_2	-3.60	-4.74	-5.91	-2.69
school variance	u_j	~N(0, 34.71)	~N(0, 41.97)	~N(0, 45.63)	~N(0, 35.36)
student (residual) variance	e_{ij}	~N(0, 51.00)	~N(0, 63.39)	~N(0, 59.81)	~N(0, 60.62)

Tests for Division (fixed) effects. The null hypothesis for this question is that age (modeled using Division) has no effect on student scores. As mentioned earlier, the likelihood ratio test for fixed effects requires that the two models in comparison be fitted using full Maximum Likelihood computations. Chi-square values, degrees of freedom and p-

values obtained from the likelihood ratio test after refitting the original models and fitting the nested models are as follows.

Likelihood Ratio Test for Fixed Effects (Division)			
Year	**Test**	**dof**	**p-value**
2006	X^2=2.0702	2	.3552
2007	X^2=4.2024	2	.1223
2008	X^2=10.867	2	.0044
2009	X^2=3.3416	2	.1881

The results show no evidence for significant Division effects for 2006, 2007 or 2009 – but they do for 2008 at alpha=.05, and even at alpha=.01.

A further look at Division effects. Another way to look at Division as a factor – one that might be more intuitive – comes from Bayesian statistics, and involves examining the 95% confidence intervals of the predicted means for Elementary, Junior High, and Senior High Division levels. These means and confidence intervals are obtained using a method called "Gibbs sampling" which samples from the posterior distribution of the parameters in the model. Gibbs sampling, a special case of "Markov chain Monte Carlo sampling," simulates random samples from the probability distributions of the parameters in order to estimate the parameters themselves. It holds some of the parameters in the model fixed while sampling others, and after multiple iterations begins to converge on estimates for all the parameters in the model and their 95% confidence intervals (Bates, 2006).

Below are the Bayesian estimates for the Division means and their 95% confidence intervals after running 100,000 iterations of this technique for each year of data. The Empirical Means are computed from the actual data, the Traditional Estimates are predicted from the

regression lines, and the Bayesian Estimates are those obtained using Gibbs sampling. Notice that the Bayesian parameter estimates differ slightly from the traditionally computed estimates due to random sampling error.

2006 Means						
Division	**N**	**Empirical Mean**	**Traditional Estimate**	**Bayesian Estimate**	**95% CI Lower Limit**	**95% CI Upper Limit**
Elementary	55	79.86	79.236	79.804	77.236	82.3484
Junior	72	79.74169	77.766	78.740	72.556	84.8038
Senior	56	77.20694	75.636	76.641	70.212	83.1297

2007 Means						
Division	**N**	**Empirical Mean**	**Traditional Estimate**	**Bayesian Estimate**	**95% CI Lower Limit**	**95% CI Upper Limit**
Elementary	66	81.63	81.183	81.646	78.887	84.398
Junior	106	76.60154	76.571	76.219	69.701	82.7468
Senior	61	78.12828	76.440	77.174	70.047	84.2668

2008 Means						
Division	**N**	**Empirical Mean**	**Traditional Estimate**	**Bayesian Estimate**	**95% CI Lower Limit**	**95% CI Upper Limit**
Elementary	77	79.61	80.398	80.116	77.782	82.5266
Junior	86	76.46658	74.213	75.399	69.68	81.2561
Senior	54	78.60403	74.483	76.499	69.589	82.1106

2009 Means						
Division	**N**	**Empirical Mean**	**Traditional Estimate**	**Bayesian Estimate**	**95% CI Lower Limit**	**95% CI Upper Limit**
Elementary	113	79.27	80.306	79.6992	77.744	81.6053
Junior	92	78.17173	77.054	77.6156	72.627	82.5749
Senior	55	81.27160	77.612	79.6147	73.683	85.4403

Examination of the confidence intervals for 2006 and 2009 indicate a lack of sufficient evidence to reject the null hypothesis that Division has no effect on the average score for each student; for both years, confidence intervals for all three division levels contain each other's midpoints.

For 2007 and 2008 however, this is not the case. In the 2007 data, the midpoints for the Junior and Senior confidence intervals are respectively 76.22 and 77.16 – neither of which are contained in the Elementary division's confidence interval. In the data for 2008, the Junior and Senior midpoints are 75.47 and 75.85, which also are not contained in the confidence interval for the Elementary division of that year.

So as confidence intervals overlap for 2007's and 2008's divisions, yet do not all contain each other's centers, a numerical test for significant differences is desirable for clarity. Such a test is brought to light by Wolfe and Hanley (2002), where the distance between confidence interval midpoints is compared to functions of the standard errors (half the length and divided by 1.96) of the confidence intervals.

$$2\left(SE_A^{\,2}+SE_B^{\,2}\right)^{\frac{1}{2}} \;<\; \text{midpoint}_B-\text{midpoint}_A \;<\; 2SE_A+2SE_B$$

These are the calculated values used in the logical tests for 2007:

Elementary vs Junior	7.23 <? 5.42 <? 9.47
Elementary vs Senior	7.78 <? 4.49 <? 10.07
Junior vs Senior	9.85 <? .933 <? 13.91

and for 2008:

Elementary vs Junior	6.38 <? 4.69 <? 8.33
Elementary vs Senior	6.83 <? 4.30 <? 8.81
Junior vs Senior	8.70 <? .38 <? 12.29

Because the test does not hold for either year, the claim cannot be made that Bayesian confidence intervals show evidence for significant differences.

Proportion of variance due to school grouping. The **"Inter-class Correlation"** is an estimate of the total proportion of variance in the model that can be attributed to school grouping:

$$ICC = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u + \hat{\sigma}^2_e}$$

This gives researchers an excellent feel for how much a random factor impacts the outcome variable – or at least how much it impacts the overall variance of the outcome variable. The Inter-class Correlations are easily computed, as the required variance estimates were reported earlier in this paper.

Inter-class Correlations			
2006	**2007**	**2008**	**2009**
.405	.398	.433	.368

All years of data report that approximately 40% of the total variance in the model is attributed to the variance between schools. This appears to be a very large portion of the

total variance, and a test for the significance of it will be helpful.

Tests for School (random) effects. Testing for significance of random effects, or the effect of schools, is available using the RLRsim ("Exact (Restricted) Likelihood Ratio tests for mixed and additive models") package for R. Because the likelihood ratio test statistic itself follows a distribution, the RLRsim package samples from the likelihood ratio's "exact finite sampling distribution" to obtain an observed value (Scheipl, 2010). This observed value is then compared with a chi-square distribution to calculate a p-value.

The null hypothesis for this test is that the variance attributed to the random component $\mathbf{u_j}$ is equal to zero. Restricted likelihood ratio test values and p-values are given below from a sample of 100,000 iterations of the process using RLRsim.

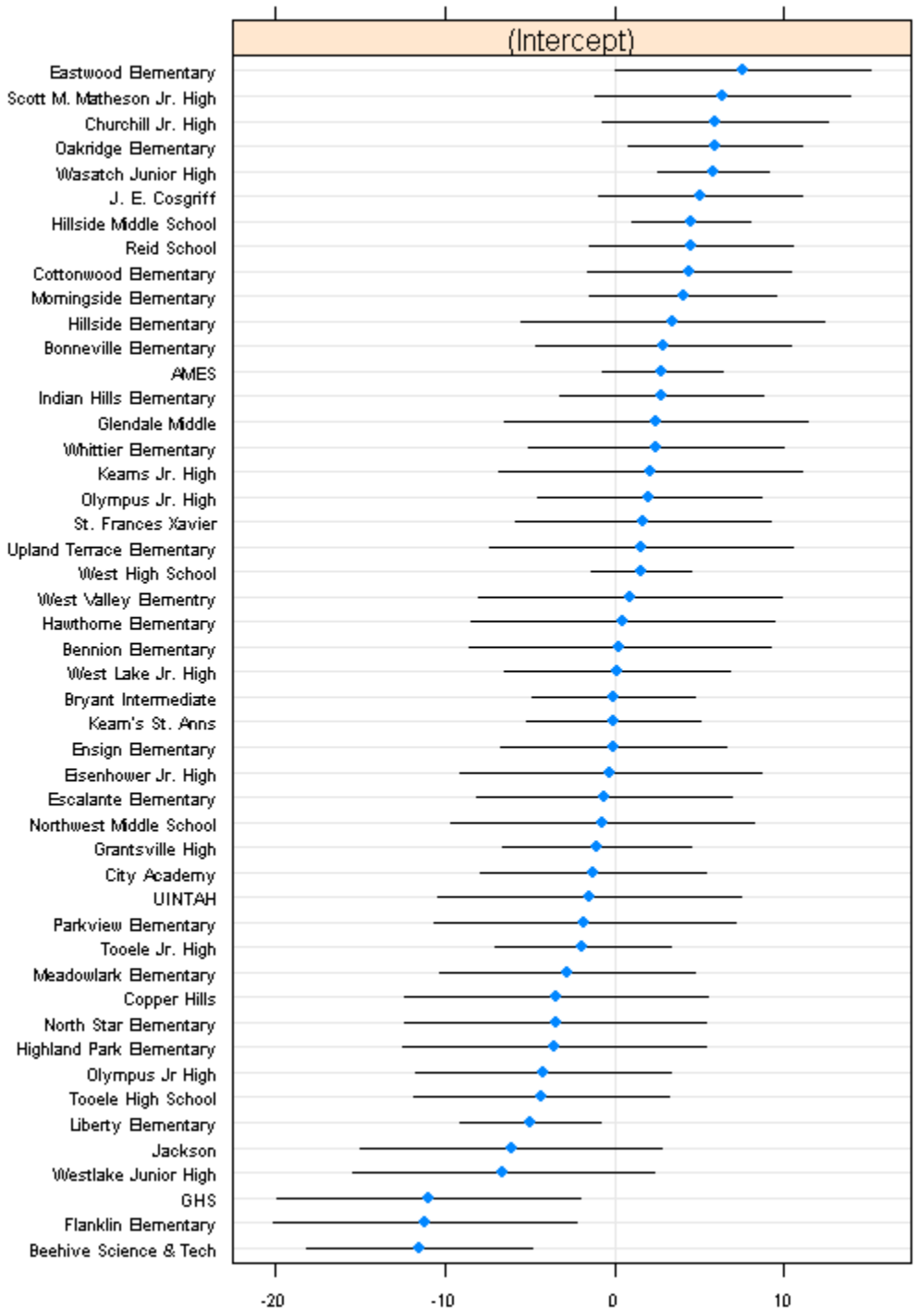
Random Effects p-values		
Year	**Test**	**p-value**
2006	RLRT=19.41	$1 \cdot 10^{-5}$
2007	RLRT=46.23	$2.2 \cdot 10^{-16}$
2008	RLRT=36.38	$2.2 \cdot 10^{-16}$
2009	RLRT=30.89	$2.2 \cdot 10^{-16}$

The p-values indicate strong evidence for random effects, which agrees with the Inter-class Correlations' report that a large proportion of the variance in the model for each year is due to variance between schools.

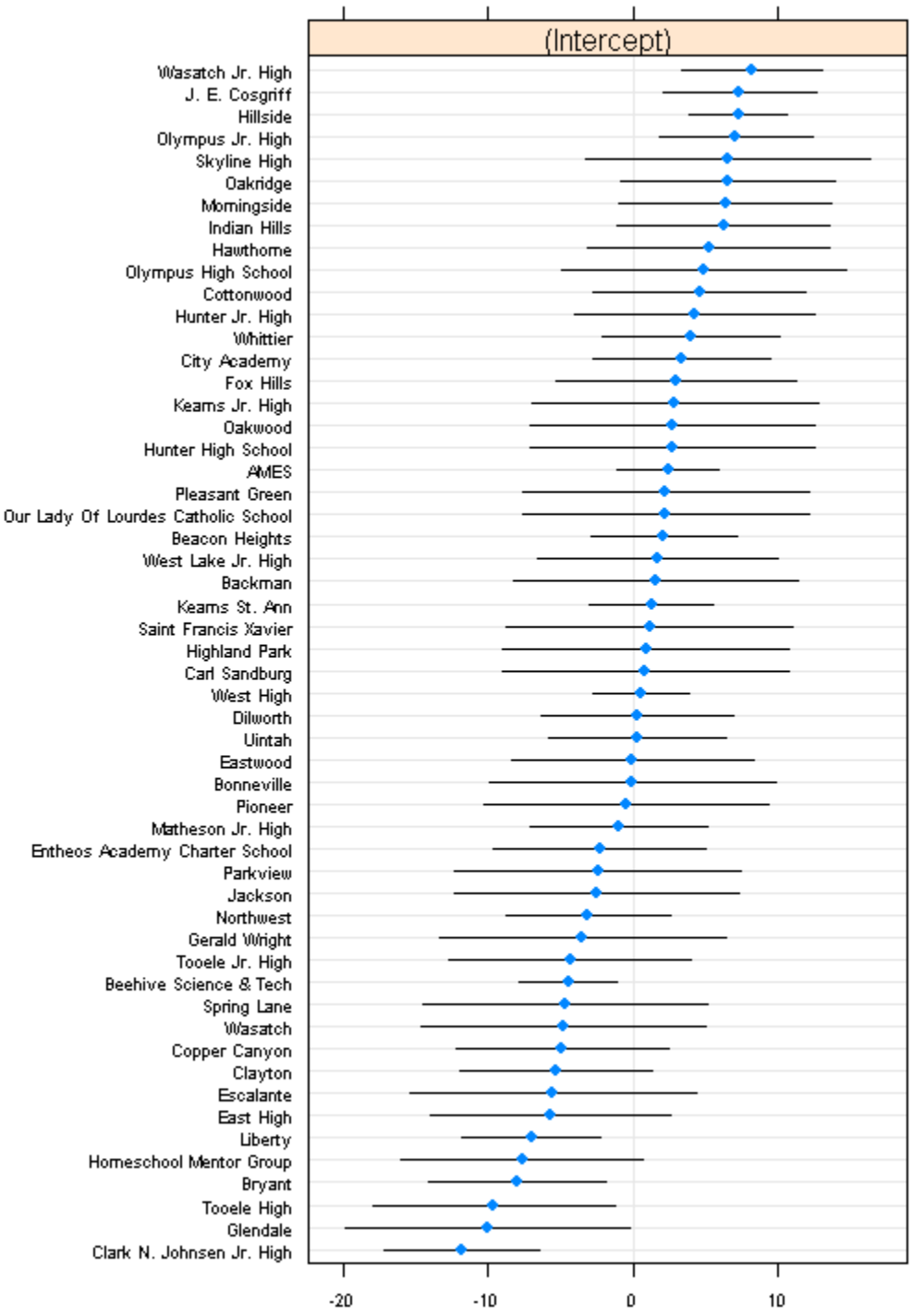
Estimates of individual school contributions. As the variance of $\mathbf{u_j}$ has been estimated for each year, estimates of $\mathbf{u_j}$ itself for every "j" (school) are displayed in a graph below. These are the estimated contributions, or adjustments, to the outcome score perpetuated by

each school. 95% confidence intervals are shown as well.

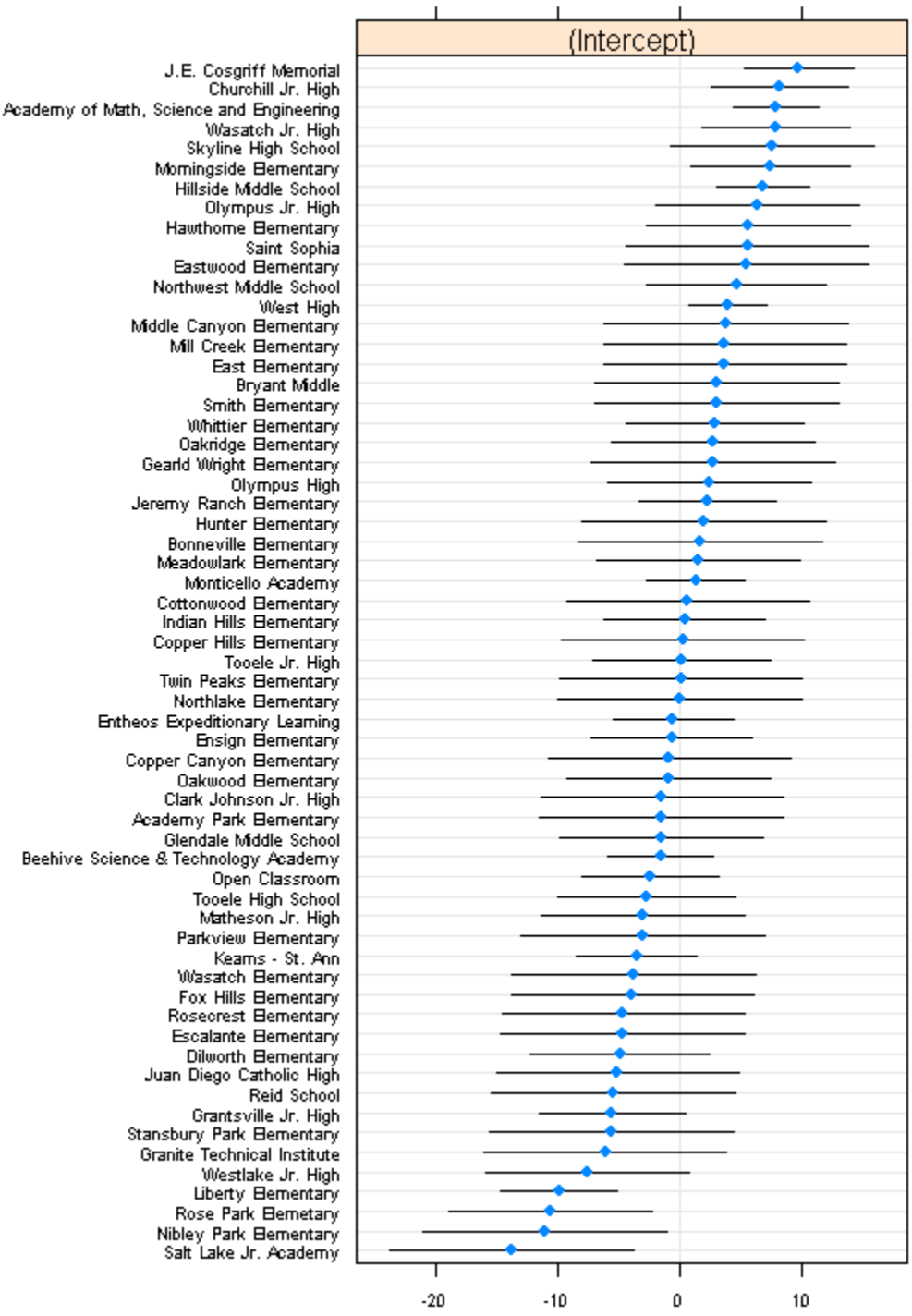
2006


(Intercept)
Eastwood Elementary
Scott M. Matheson Jr. High
Churchill Jr. High
Oakridge Elementary
Wasatch Junior High
J. E. Cosgriff
Hillside Middle School
Reid School
Cottonwood Elementary
Morningside Elementary
Hillside Elementary
Bonneville Elementary
AMES
Indian Hills Elementary
Glendale Middle
Whittier Elementary
Kearns Jr. High
Olympus Jr. High
St. Frances Xavier
Upland Terrace Elementary
West High School
West Valley Elementry
Hawthorne Elementary
Bennion Elementary
West Lake Jr. High
Bryant Intermediate
Kearn's St. Anns
Ensign Elementary
Eisenhower Jr. High
Escalante Elementary
Northwest Middle School
Grantsville High
City Academy
UINTAH
Parkview Elementary
Tooele Jr. High
Meadowlark Elementary
Copper Hills
North Star Elementary
Highland Park Elementary
Olympus Jr High
Tooele High School
Liberty Elementary
Jackson
Westlake Junior High
GHS
Flanklin Elementary
Beehive Science & Tech
-20
-10
0
10

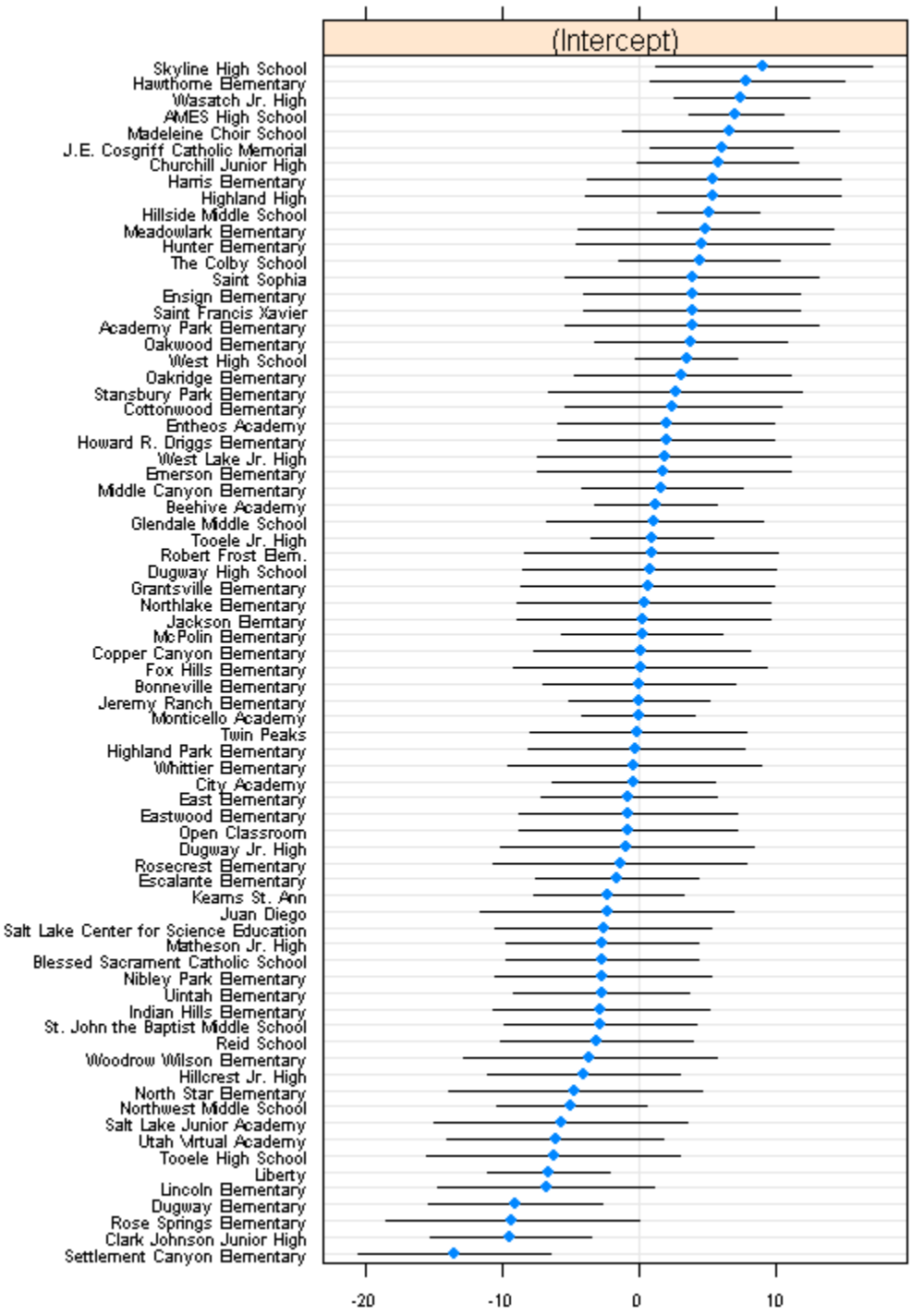
2007


(Intercept)
Wasatch Jr. High
J. E. Cosgriff
Hillside
Olympus Jr. High
Skyline High
Oakridge
Morningside
Indian Hills
Hawthorne
Olympus High School
Cottonwood
Hunter Jr. High
Whittier
City Academy
Fox Hills
Kearns Jr. High
Oakwood
Hunter High School
AMES
Pleasant Green
Our Lady Of Lourdes Catholic School
Beacon Heights
West Lake Jr. High
Backman
Kearns St. Ann
Saint Francis Xavier
Highland Park
Carl Sandburg
West High
Dilworth
Uintah
Eastwood
Bonneville
Pioneer
Matheson Jr. High
Entheos Academy Charter School
Parkview
Jackson
Northwest
Gerald Wright
Tooele Jr. High
Beehive Science & Tech
Spring Lane
Wasatch
Copper Canyon
Clayton
Escalante
East High
Liberty
Homeschool Mentor Group
Bryant
Tooele High
Glendale
Clark N. Johnsen Jr. High
-20
-10
0
10

2008


(Intercept)
J.E. Cosgriff Memorial
Churchill Jr. High
Academy of Math, Science and Engineering
Wasatch Jr. High
Skyline High School
Morningside Elementary
Hillside Middle School
Olympus Jr. High
Hawthorne Elementary
Saint Sophia
Eastwood Elementary
Northwest Middle School
West High
Middle Canyon Elementary
Mill Creek Elementary
East Elementary
Bryant Middle
Smith Elementary
Whittier Elementary
Oakridge Elementary
Gearld Wright Elementary
Olympus High
Jeremy Ranch Elementary
Hunter Elementary
Bonneville Elementary
Meadowlark Elementary
Monticello Academy
Cottonwood Elementary
Indian Hills Elementary
Copper Hills Elementary
Tooele Jr. High
Twin Peaks Elementary
Northlake Elementary
Entheos Expeditionary Learning
Ensign Elementary
Copper Canyon Elementary
Oakwood Elementary
Clark Johnson Jr. High
Academy Park Elementary
Glendale Middle School
Beehive Science & Technology Academy
Open Classroom
Tooele High School
Matheson Jr. High
Parkview Elementary
Kearns - St. Ann
Wasatch Elementary
Fox Hills Elementary
Rosecrest Elementary
Escalante Elementary
Dilworth Elementary
Juan Diego Catholic High
Reid School
Grantsville Jr. High
Stansbury Park Elementary
Granite Technical Institute
Westlake Jr. High
Liberty Elementary
Rose Park Elementary
Nibley Park Elementary
Salt Lake Jr. Academy
-20
-10
0
10

2009


(Intercept)
Skyline High School
Hawthorne Elementary
Wasatch Jr. High
AMES High School
Madeleine Choir School
J.E. Cosgriff Catholic Memorial
Churchill Junior High
Harris Elementary
Highland High
Hillside Middle School
Meadowlark Elementary
Hunter Elementary
The Colby School
Saint Sophia
Ensign Elementary
Saint Francis Xavier
Academy Park Elementary
Oakwood Elementary
West High School
Oakridge Elementary
Stansbury Park Elementary
Cottonwood Elementary
Entheos Academy
Howard R. Driggs Elementary
West Lake Jr. High
Emerson Elementary
Middle Canyon Elementary
Beehive Academy
Glendale Middle School
Tooele Jr. High
Robert Frost Elem.
Dugway High School
Grantsville Elementary
Northlake Elementary
Jackson Elementary
McPolin Elementary
Copper Canyon Elementary
Fox Hills Elementary
Bonneville Elementary
Jeremy Ranch Elementary
Monticello Academy
Twin Peaks
Highland Park Elementary
Whittier Elementary
City Academy
East Elementary
Eastwood Elementary
Open Classroom
Dugway Jr. High
Rosecrest Elementary
Escalante Elementary
Kearns St. Ann
Juan Diego
Salt Lake Center for Science Education
Matheson Jr. High
Blessed Sacrament Catholic School
Nibley Park Elementary
Uintah Elementary
Indian Hills Elementary
St. John the Baptist Middle School
Reid School
Woodrow Wilson Elementary
Hillcrest Jr. High
North Star Elementary
Northwest Middle School
Salt Lake Junior Academy
Utah Virtual Academy
Tooele High School
Liberty
Lincoln Elementary
Dugway Elementary
Rose Springs Elementary
Clark Johnson Junior High
Settlement Canyon Elementary
-20
-10
0
10

Notice that the estimated values and confidence intervals for schools are ordered from those most positively affecting the outcome to those most negatively effecting the outcome, from top to bottom. For every year, the highest 95% confidence interval does not overlap the lowest at all. This agrees with the conclusion that the effect of schools is significant.

Judge Deviation Data

Higher scoring judges. First we consider judges who score higher, on average, than the average scores from other judges of the same projects. The highest ten will be used from each year.

2006 Judges with Higher Mean Deviations					
Judge	**Mean Deviation**	**Years XP in Field**	**Previous Judging XP**	**Degree**	**Field**
119	17.07	19	No	BS	Chemical Engineering
88	13.91	---	No	BS	Zoology
32	11.90	9	Yes	BS	Engineering
79	11.49	10	No	PhD	Molecular Biology, Genetics
30	11.45	4	No	PhD	Chemistry, Biology, Medicine
59	10.39	8	Yes	MS	Engineering
43	9.97	11	Yes	BS	Civil Engineering
27	9.49	12	No	PhD	Pharmacology, Immunology
122	8.72	7	Yes	BS	Civil Engineering
69	8.39	40	No	BS	Mechanical Engineering

2007 Judges with Higher Mean Deviations				
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Field**
124	11.19	7	PhD	Optics, Physics
122	10.33	7	BS	Nuclear Engineering
85	9.91	1	MS	Engineering, Physics, Chemistry
83	9.39	10	BS	Microbiology, Zoology
102	9.31	6	PhD	Engineering, Physics
134	9.19	24	MS	Engineering, Physics
24	8.58	16	PhD	Microbiology, Medicine
34	8.41	35	MS	Microbiology, Medicine
28	8.25	12	PhD	Computer Science, Mathematics
84	8.20	11	BS	Structural Engineering

2008 Judges with Higher Mean Deviations							
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Previous Judging XP**	**City**	**Company**	**Field**
175	17.54	41	BS	No	SLC	Williams	Drafting
177	14.42	4	PhD	No	SLC	U of U	Finance, Physics
36	13.51	7	BS	Yes	SLC	Myriad Genetics	Zoology
155	12.35	0	PhD	No	---	U of U	Chemical, Bioengineering
62	9.91	13	PhD	Yes	SLC	U of U	Physics
91	9.43	4	BS	Yes	Holladay	U of U	Engineering, Medicine
138	9.38	2	MS	No	Bountiful	---	Behavioral Science
20	9.25	15	PhD	Yes	SLC	U of U	Biochemistry, Chemistry
74	9.14	6	BS	Yes	SLC	Idaho Technology	Molecular & Microbiology
96	9.08	4	MS	No	SLC	U of U	Behavioral Science

2009 Judges with Higher Mean Deviations							
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Previous Judging XP**	**City**	**Company**	**Field**
117	16.65	7	PhD	Yes	SLC	U of U	Bioengineering
63	15.11	22	PhD	Yes	SLC	---	Computers
49	14.32	0	MS	No	SLC	---	Mechanical Engineering
108	14.00	8	MS	Yes	SLC	Utah Governor's Office	Atmospheric Science
62	12.31	10	MS	Yes	SLC	U of U	Physics, Engineering
57	12.28	25	MS	Yes	Bountiful	Corps of Engineers	Engineering
99	11.92	40	BS	Yes	SLC	Williams	Drafting
148	11.48	8	BS	Yes	SLC	Myriad Genetics	Zoology
237	10.13	>10	PhD	Yes	SLC	Energy Solutions	Environment
206	10.01	1	BS	No	SLC	Salt Lake School District	Neuroscience

The tables appear to show a trend with judges who are involved in engineering or biology. This may be partially due to the fact that many participating science fair judges are engineers or biologists, but there does seem to be a greater concentration in these tables nonetheless. Other than that, though, there are no noticeable patterns in the level of education (which agrees with the statistical results), whether or not the judge has judged

previously, years of experience, or the nature of the company for whom the judge works.

Lower scoring judges. Next we consider the judges who score lower, on average, than the average scores from other judges of the same projects. The lowest ten will be used from each year.

2006 Judges with Lower Mean Deviations					
Judge	**Mean Deviation**	**Years XP in Field**	**Previous Judging XP**	**Degree**	**Field**
61	-14.57	6	No	BS	Computer Science
24	-12.83	9	Yes	BS	Civil Engineering
63	-11.88	1	No	MS	Mechanical Engineering
42	-11.03	21	No	PhD	Bioengineering, Chemistry
108	-9.92	20	Yes	PhD	Microbiology
106	-9.88	4	No	MS	Geology, Geophysics
11	-9.75	40	Yes	BA	Behavioral Science
78	-9.55	21	No	BS	Manufacturing Engineering
5	-9.41	4	Yes	MS	Meteorology
84	-8.68	0	No	BS	Meteorology, Earth Science

2007 Judges with Lower Mean Deviations				
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Field**
66	-17.42	1	PhD	Biology, Chemistry, Pharmacology
144	-12.11	>10	PhD	Computer Science
126	-11.94	2	BS	Behavioral Science
13	-11.61	40	MS	Physics, Environment, Medicine
204	-10.88	---	---	Chemistry
140	-9.99	3	BS	Environment, Biology
132	-9.85	9	BS	Mathematics
55	-9.26	17	PhD	Biomedical Physics
31	-9.01	25	PhD	Electrical Engineering, Mathematics
146	-8.99	---	PhD	Biology

2008 Judges with Lower Mean Deviations							
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Previous Judging XP**	**City**	**Company**	**Field**
168	-21.68	5	PhD	No	SLC	U of U	Genetics
160	-18.56	14	MS	Yes	Tooele	Department of Defense	Microbiology
6	-13.00	29	BS	Yes	SLC	U of U	Computers, Physics
57	-12.60	12	PhD	Yes	SLC	American Chemical Society	Organic Chemistry
134	-10.81	16	PhD	Yes	SLC	U of U	Statistics, Genetics
174	-10.05	1	MS	No	Taylorsville	Kleinfelder	Meteorology
166	-10.02	2	MS	Yes	---	Kearns Jr. High	Biology, Chemistry
67	-9.27	30	PhD	Yes	Dugway	Dugway Proving Grounds	Physics
128	-8.96	29	PhD	Yes	Clearfield	Northrop Grumman	Computers
3	-8.33	13	BS	Yes	SLC	Colvin Engineering	Engineering

2009 Judges with Lower Mean Deviations							
Judge	**Mean Deviation**	**Years XP in Field**	**Degree**	**Previous Judging XP**	**City**	**Company**	**Field**
116	-20.89	6	PhD	Yes	SLC	U of U	Engineering
54	-17.97	34	BS	Yes	SLC	Hercules ATK	Chemical Engineering
21	-15.78	4	PhD	Yes	SLC	U of U	Biochemistry
26	-14.22	25	PhD	Yes	SLC	U of U	Computers
33	-13.89	30	PhD	Yes	SLC	U of U	Computers
114	-13.74	9	MS	Yes	Murray	Nolte Associates	Engineering
104	-13.00	38	BS	Yes	West Jordan	Motorola	Engineering
155	-11.71	---	MS	Yes	SLC	Utah Business Accelerator	Business
14	-11.56	4	MS	Yes	SLC	L-3	Computers, Physics
82	-10.69	16	PhD	Yes	SLC	U of U	Molecular Biology

The same concentration of engineers and biologists emerges in these lower scoring tables as it did in the higher scoring tables. In 2008 and 2009, though, almost all of the lowest scoring judges had previous judging experience – slightly more so than the highest scoring judges. Again, there are not enough repeating companies listed to make any kind of inferences about the judges' places of work.

Average deviations for more general categories of judges. It may be interesting to look at the average of all judges' average deviations using a broader classification scheme. Recall that this is the original classification scheme which judges self-report:

behavioral science
biochemistry
botany
chemistry
computer science
earth science
engineering
environmental management
mathematics
medicine
microbiology
physics
space science
zoology

Here are the more general groups into which the author has grouped judges for the upcoming tables in order that more judges can be counted in each group:

behavioral science (behavioral science)
chemistry & microbiology (biochemistry, chemistry, microbiology)
earth science (earth science, environmental management)
macro-biology (botany, medicine, zoology)
mathematics (computer science, engineering, mathematics, physics, space science)

Note that because most judges are self-reported into several categories, almost every judge is in more than one of the original categories, and thus many are even in more than one of these more general groups. Thus, average deviation scores are correlated between groups – but this can be viewed as an accurate reflection of the way judges self-report.

When viewing judges with lower and higher average score deviations in the previous tables, the average deviation was calculated for every judge, and individual judges were examined. In the following tables, the judges' average deviation scores were averaged for every judge within each of the more general groups:

2006 Group Mean Deviations		
N (correlated)	**Group of Judges**	**Mean of Mean Deviations**
16	behavioral science	-.193
50	chemistry & microbiology	1.055
53	earth science	-1.564
47	macro-biology	.336
120	mathematics	-.132

2007 Group Mean Deviations		
N (correlated)	**Group of Judges**	**Mean of Mean Deviations**
2	behavioral science	-5.320
48	chemistry & microbiology	1.002
35	earth science	-.586
38	macro-biology	.881
103	mathematics	.099

2008 Group Mean Deviations		
N (correlated)	**Group of Judges**	**Mean of Mean Deviations**
7	behavioral science	3.378
50	chemistry & microbiology	-.901
24	earth science	-1.341
37	macro-biology	1.244
87	mathematics	.163

2009 Group Mean Deviations		
N (correlated)	**Group of Judges**	**Mean of Mean Deviations**
22	behavioral science	-.202
61	chemistry & microbiology	-.165
39	earth science	2.406
54	macro-biology	1.662
99	mathematics	-.389

Much disparity exists between the numbers of judges in each group for each year. Still, for most groups the average of mean deviations is between -2 and 2. There seems to be no pattern for any of the groups across any years, and there are no noticeably divergent results within any one year. The exceptions, of course, are the groups with low numbers of judges, where measurements of the mean are expected to deviate more drastically. The "macro-biology" group (botany, medicine, and zoology) is the only group whose mean is always positive, but even those means are not substantially higher than others, nor are they very consistent.

Discussion

Not finding evidence for judging bias is basically good. If there is not enough statistical evidence to conclude judging bias exists for a particular question, we fail to reject the null hypothesis that "there is no bias." However, this does not ensure that there is no bias. After all, *failing to reject* a null hypothesis is not the same as *accepting* that null hypothesis (Dallal, 2007). In other words, if we do not possess evidence of judging bias, this does not necessarily mean that we *do* possess positive evidence that no bias exists. So essentially, the best possible outcome for any of the study questions would be to conclude that there is a lack of evidence for judging bias. To conclude otherwise would of course insinuate that scores given by judges are dependent on a factor that is unrelated to the quality of the science fair project.

With this in mind, most of the outcomes of this study were favorable. None of the judges seemed to judge higher or lower based on their own education or area of expertise. No distinct patterns could be discerned among judges who scored lower or higher than the averages for the projects they judged. And where the students' age, or division, was the independent variable, bias was only shown positively for 2008.

So what does that mean – that judges were only giving scores as a function of age in 2008? Was there something that happened to influence judging dynamics that year, or was it simply Type I error? The author did, after all, use more than one test of significance for division effects without adjusting for Type I error – in addition to the fact that the other test – namely, the Bayesian confidence intervals – did not show significance for 2008.

With so much of the variance in the linear model being accounted for by schools, it is natural

to wonder what the 'mechanism' or reason is for this variance. The scores awarded to projects within schools may differ for several reasons, and it is difficult to decide if any of these reasons are accurate. Some schools may have a systemized approach to preparing students for science fairs, and may be very good at it. Some schools probably have superior methods for teaching science and technology. The students in some schools may have better support systems in their lives due to better socioeconomic status or family situations. From another angle, judges can have bias toward or against schools. For example, in 2006 “Beehive Science & Tech” was ironically at the bottom of the list of the model's school contributions. Could it be that judges expected more because of the name of the school and judged more harshly? Again, the accuracy of any of these ideas would be difficult to determine conclusively.

Possible study limitations. For the judges who tended to score higher or lower than the average scores of the projects they judged, perhaps more solid conclusions could be drawn if more information about judges were available. Deeper facts about a judge, such as age, quality of life, free time given to scientific endeavors, or family life might serve as better data for finding general patterns of higher or lower scoring. Of course, the case could simply be that a judge's propensity for scoring higher or lower is not derived from anything chronic: judges, being people, have good days and bad days. Personality differences could also come into play: one judge having a bad day might give way to an angry disposition while another judge having a bad day may feel softened toward the strangers around him who are not part of his personal problems. Finally, it is also possible that deviations among scores are simply due to random error. But whatever the reason for some judges' scoring higher or lower, a greater amount of information about the judges would give researchers a better chance at identifying what that reason is or what it is not.

Another limitation is on the part of the author: as just mentioned, Type I error was not

adjusted for, even though two tests were used to test for Division effects in the hierarchical model.

One question that may not be obvious at first is 'Are “areas of expertise” the same as “areas of interest?”' While this was not addressed in the main body of this study, the author has reflected on whether equating these two concepts has been reasonable.

Finally, as the name of this paper is “Science Fair Fairness,” due diligence would require the admission that the statistical tests involved have not been comprehensive in their search for 'fairness'. Specifically, these tests measure *systematic* sources of bias: those that are identifiable and shift the mean of a group of scores in a consistent direction, for example. Random error is that which cannot be attributed to anything specific, such as the difference of scores from one judge to another on any given project, or the way one judge judges differently on one day than another day. Random error would not affect the mean of a group of scores in the same direction each time the same population is sampled. In other words, if random error were the only source of error, the sample mean would literally oscillate 'randomly' about the population mean. If this were the case with the Salt Lake Valley Science and Engineering Fair, this study would never find statistical bias in any of the statistical tests that have been used in this paper – but scores would always be affected by the random measurement error that is part of subjective judgment.

That having been said, a generalizability study would be an interesting new direction for this backdrop. One could aim to discover the optimal number of judges to assign to a single student project, for example.

Appendix

Descriptive Statistics for Age Data

Below are the mean scores given to every school for each year – that is, the mean of all scores given to all student projects within each school. Where "NA" is reported in the Standard Deviation column, there was only one student competing from that school for that year.

2006		
School	**Mean Score**	**Standard Deviation**
AMES	78.75	8.52
Beehive Science & Tech	60.64	0.66
Bennion Elementary	80	NA
Bonneville Elementary	84.25	4.6
Bryant Intermediate	77.71	7.11
Churchill Jr. High	84.52	6.9
City Academy	73.75	6.55
Copper Hills	70.75	NA
Cottonwood Elementary	85.3	8.44
Eastwood Elementary	92.32	3.08
Eisenhower Jr. High	77.2	NA
Ensign Elementary	79.12	9.81
Escalante Elementary	78.28	3.85
Franklin Elementary	51.67	NA
GHS	50.67	NA
Glendale Middle	83.75	NA
Grantsville High	74.29	6.01
Hawthorne Elementary	80.4	NA
Highland Park Elementary	70.5	NA
Hillside Elementary	87.78	NA
Hillside Middle School	82.74	4.95
Indian Hills Elementary	83	5.28
J. E. Cosgriff	84.74	3.72
Jackson	64.2	NA
Kearn's St. Anns	77.7	9.64
Kearns Jr. High	83	NA

2006		
School	**Mean Score**	**Standard Deviation**
Liberty Elementary	73.6	5.6
Meadowlark Elementary	74.43	6.74
Morningside Elementary	84.49	3.9
North Star Elementary	70.67	NA
Northwest Middle School	76	NA
Oakridge Elementary	86.6	4.23
Olympus Jr High	69.5	1.41
Olympus Jr. High	79.41	9.4
Parkview Elementary	74.9	NA
Reid School	83.78	4.68
Scott M. Matheson Jr. High	88.75	6.25
St. Frances Xavier	80.7	0.99
Tooele High School	68.21	4.89
Tooele Jr. High	75.46	9.6
UINTAH	75.57	NA
Upland Terrace Elementary	83.17	NA
Wasatch Junior High	83.89	6.31
West High School	77.51	8.86
West Lake Jr. High	75.84	6.2
West Valley Elementary	81.57	NA
Westlake Junior High	61.6	NA
Whittier Elementary	83.4	8.49

2007		
School	**Mean Score**	**Standard Deviation**
AMES	79.02	7.36
Backman	85	NA
Beacon Heights	83.69	4.33
Beehive Science & Tech	71.72	9.74
Bonneville	81	NA
Bryant	66.17	10.33
Carl Sandburg	83.25	NA
City Academy	80.72	2.87
Clark N. Johnsen Jr. High	62.24	11.28
Clayton	69.19	9.71
Copper Canyon	73.75	3.25
Cottonwood	88.05	4.16
Dilworth	81.53	2.83

2007		
School	**Mean Score**	**Standard Deviation**
East High	66.38	17.5
Eastwood	81.08	0.95
Entheos Academy Charter	74.61	11.1
Escalante	67.25	NA
Fox Hills	86.33	14.38
Gerald Wright	72.33	NA
Glendale	51.25	NA
Hawthorne	90.3	3.25
Highland Park	83.33	NA
Hillside	84.34	5.24
Homeschool Mentor Group	63.08	9.78
Hunter High School	83.17	NA
Hunter Jr. High	84	0.71
Indian Hills	90.52	5.28
J. E. Cosgriff	85.5	4.28
Jackson	74.75	NA
Kearns Jr. High	83.8	NA
Kearns St. Ann	77.96	9.64
Liberty	72.95	12.33
Matheson Jr. High	75.25	3.47
Morningside	90.75	7.17
Northwest	72.63	5.91
Oakridge	90.97	2.08
Oakwood	88	NA
Olympus High School	88.63	NA
Olympus Jr. High	85.13	4.65
Our Lady Of Lourdes Catholic	82.17	NA
Parkview	75	NA
Pioneer	80	NA
Pleasant Green	86.8	NA
Saint Francis Xavier	79.33	NA
Skyline High	92.86	NA
Spring Lane	69.4	NA
Tooele High	59.55	12.56
Tooele Jr. High	68.88	5.13
Uintah	81.51	11.68
Wasatch	69	NA
Wasatch Jr. High	86.01	5.06
West High	76.98	8.52
West Lake Jr. High	79.42	2.24

2007		
School	**Mean Score**	**Standard Deviation**
Whittier	86.3	4.57

2008		
School	**Mean Score**	**Standard Deviation**
AMES	82.99	8.63
Academy Park Elementary	77	NA
Beehive Science & Tech	72.55	6
Bonneville Elementary	84.4	NA
Bryant Middle	81.2	NA
Churchill Jr. High	84.25	3.68
Clark Johnson Jr. High	70.88	NA
Copper Canyon Elementary	78.5	NA
Copper Hills Elementary	81	NA
Cottonwood Elementary	82	NA
Dilworth Elementary	73.39	12.01
East Elementary	89	NA
Eastwood Elementary	93	NA
Ensign Elementary	79.57	4.05
Entheos Expeditionary Learning	76.68	9.7
Escalante Elementary	69.5	NA
Fox Hills Elementary	71.5	NA
Gerald Wright Elementary	86.8	NA
Glendale Middle School	71.75	2.24
Granite Technical Institute	60.4	NA
Grantsville Jr. High	67.26	6.35
Hawthorne Elementary	89.63	1.94
Hillside Middle School	81.6	6.9
Hunter Elementary	85	NA
Indian Hills Elementary	80.98	3.41
J.E. Cosgriff Memorial	85.27	7.12
Jeremy Ranch Elementary	83.1	3.39
Juan Diego Catholic High	62.75	NA
Kearns - St. Ann	70.09	7.03
Liberty Elementary	69.07	6.81
Matheson Jr. High	69.44	1.5
Meadowlark Elementary	82.93	10.85
Middle Canyon Elementary	89.25	NA
Mill Creek Elementary	89	NA

2008		
School	**Mean Score**	**Standard Deviation**
Monticello Academy	78.51	7.6
Morningside Elementary	90.27	9.48
Nibley Park Elementary	55	NA
Northlake Elementary	80.5	NA
Northwest Middle School	80.91	6.19
Oakridge Elementary	85.03	6.75
Oakwood Elementary	78.96	10.31
Olympus High	78.47	2.64
Olympus Jr. High	84.73	14.32
Open Classroom	73.32	13.02
Parkview Elementary	73.4	NA
Reid School	61.71	NA
Rose Park Elementary	62.88	10.78
Rosecrest Elementary	69.75	NA
Saint Sophia	93.25	NA
Salt Lake Jr. Academy	42.33	NA
Skyline High School	87.06	6.14
Smith Elementary	87.33	NA
Stansbury Park Elementary	67.6	NA
Tooele High School	70.64	9.65
Tooele Jr. High	74.41	5.72
Twin Peaks Elementary	80.67	NA
Wasatch Elementary	71.6	NA
Wasatch Jr. High	84.22	3.93
West High	78.68	9.57
Westlake Jr. High	61.91	7.55
Whittier Elementary	84.56	6.78

2009		
School	**Mean Score**	**Standard Deviation**
Academy Park Elementary	90.67	NA
AMES	85.32	5.74
Beehive Academy	78.47	6.77
Blessed Sacrament	72.78	15.8
Bonneville Elementary	80.33	6.96
Churchill Junior High	85.19	8.65
City Academy	76.91	5.37
Clark Johnson Junior High	64.34	9.52

2009		
School	**Mean Score**	**Standard Deviation**
Copper Canyon Elementary	80.53	3.85
Cottonwood Elementary	84.8	2.55
Dugway Elementary	67.33	11.72
Dugway High School	79	NA
Dugway Jr. High	74.5	NA
East Elementary	79.16	11.72
Eastwood Elementary	78.75	13.79
Emerson Elementary	85	NA
Ensign Elementary	87.48	1.03
Entheos Academy	83.98	10.22
Escalante Elementary	78.07	4.41
Fox Hills Elementary	80.5	NA
Glendale Middle School	79.09	5.25
Grantsville Elementary	82	NA
Harris Elementary	95	NA
Hawthorne Elementary	92.67	2.63
Highland High	92.17	NA
Highland Park Elementary	79.78	1.45
Hillcrest Jr. High	71.23	4.1
Hillside Middle School	82.71	5.24
Howard R. Driggs Elementary	83.93	1.87
Hunter Elementary	92.75	NA
Indian Hills Elementary	75.05	0.78
J.E. Cosgriff Catholic Memorial	84.51	7.65
Jackson Elementary	81	NA
Jeremy Ranch Elementary	80.27	6.01
Juan Diego	70.6	NA
Kearns St. Ann	74.14	8.83
Liberty	72.48	11.7
Lincoln Elementary	67.63	12.2
Madeleine Choir School	89.36	0.51
Matheson Jr. High	72.97	3.9
McPolin Elementary	80.59	5.89
Meadowlark Elementary	93.4	NA
Middle Canyon Elementary	82.47	6.17
Monticello Academy	77.81	7.13
Nibley Park Elementary	75.25	7.19
North Star Elementary	67.5	NA
Northlake Elementary	81.25	NA
Northwest Middle School	70.63	5.93

2009		
School	**Mean Score**	**Standard Deviation**
Oakridge Elementary	86.13	4.77
Oakwood Elementary	86.23	2.11
Open Classroom	77.1	15.7
Reid School	75.33	8.39
Robert Frost Elementary	82.6	NA
Rose Springs Elementary	55	NA
Rosecrest Elementary	76.5	NA
Saint Francis Xavier	85.81	9.23
Saint Sophia	90.8	NA
SL Center for Science Ed.	73.73	9.72
Salt Lake Junior Academy	62	NA
Settlement Canyon Elementary	59.03	12.08
Skyline High School	94.44	0.39
St. John the Baptist	74.7	3.8
Stansbury Park Elementary	87.5	NA
The Colby School	85.56	6.92
Tooele High School	60.5	NA
Tooele Jr. High	78.08	6.41
Twin Peaks	80.05	2.05
Uintah Elementary	76.34	1.48
Utah Virtual Academy	67.5	20.51
Wasatch Jr. High	86.18	3.84
West High School	81.45	11.47
West Lake Jr. High	82.5	NA
Whittier Elementary	79.2	NA
Woodrow Wilson Elementary	70.4	NA

Post hoc Power Calculations

All power calculations are performed using an alpha of .05, since all the statistical tests in this study were performed at that level of significance. The software program "G*Power" is used to calculate effect sizes and power. Because a succinct method for calculating the power of a Kruskal-Wallis test could not be found, the effect sizes and powers of these are estimated by calculating these values as if each statistical test had been an Analysis of Variance.

Area of Expertise Tests				
Year	**Data Used**	**Test Used**	**Effect Size**	**Power**
2006	means	t-test	.1137	.1252
2007	means	t-test	.1302	.1369
2008	means	Wilcoxon	.1300	.1365
2009	means	Wilcoxon	.2022	.2678
2006	standard deviations	Wilcoxon	.2022	.2592
2007	standard deviations	t-test	-.0295	.0543
2008	standard deviations	Wilcoxon	-.0374	.0569
2009	standard deviations	Wilcoxon	-.1627	.1900

Education Level Tests				
Year	**Data Used**	**Test Used**	**Effect Size**	**Power**
2006	means	ANOVA	.1248	.1953
2007	means	ANOVA	.1053	.1636
2008	means	Kruskal-Wallis	.0973	.1415
2009	means	ANOVA	.1028	.1755
2006	standard deviations	Kruskal-Wallis	.1503	.2677
2007	standard deviations	ANOVA	.0952	.1413
2008	standard deviations	ANOVA	.0689	.0938
2009	standard deviations	Kruskal-Wallis	.1326	.2681

R Scripts

Data Import

```
Age2006 <- read.table("C:/home/project/4.R Data/Age2006.csv", header=TRUE, sep=",")
Age2007 <- read.table("C:/home/project/4.R Data/Age2007.csv", header=TRUE, sep=",")
Age2008 <- read.table("C:/home/project/4.R Data/Age2008.csv", header=TRUE, sep=",")
Age2009 <- read.table("C:/home/project/4.R Data/Age2009.csv", header=TRUE, sep=",")

AoE2006 <- read.table("C:/home/project/4.R Data/AoE2006.csv", header=TRUE, sep=",")
AoE2007 <- read.table("C:/home/project/4.R Data/AoE2007.csv", header=TRUE, sep=",")
AoE2008 <- read.table("C:/home/project/4.R Data/AoE2008.csv", header=TRUE, sep=",")
AoE2009 <- read.table("C:/home/project/4.R Data/AoE2009.csv", header=TRUE, sep=",")

Dev2006 <- read.table("C:/home/project/4.R Data/Dev2006.csv", header=TRUE, sep=",")
Dev2007 <- read.table("C:/home/project/4.R Data/Dev2007.csv", header=TRUE, sep=",")
Dev2008 <- read.table("C:/home/project/4.R Data/Dev2008.csv", header=TRUE, sep=",")
Dev2009 <- read.table("C:/home/project/4.R Data/Dev2009.csv", header=TRUE, sep=",")

Edu2006 <- read.table("C:/home/project/4.R Data/Edu2006.csv", header=TRUE, sep=",")
Edu2007 <- read.table("C:/home/project/4.R Data/Edu2007.csv", header=TRUE, sep=",")
Edu2008 <- read.table("C:/home/project/4.R Data/Edu2008.csv", header=TRUE, sep=",")
Edu2009 <- read.table("C:/home/project/4.R Data/Edu2009.csv", header=TRUE, sep=",")
```

Aggregation and other Preparation

```
# Age data - Take the mean of each student project
attach(Age2006)
m2006 <- aggregate(Age2006, by=list(school=School, project=Project, division=Division), mean)
detach(Age2006)
attach(Age2007)
m2007 <- aggregate(Age2007, by=list(school=School, project=Project, division=Division), mean)
detach(Age2007)
attach(Age2008)
m2008 <- aggregate(Age2008, by=list(school=School, project=Project, division=Division), mean)
detach(Age2008)
attach(Age2009)
m2009 <- aggregate(Age2009, by=list(school=School, project=Project, division=Division), mean)
detach(Age2009)

# Age data - Remove unneeded columns & cleanup objects
Age6m = subset(m2006, , select=c(project,school,division,Score))
Age7m = subset(m2007, , select=c(project,school,division,Score))
Age8m = subset(m2008, , select=c(project,school,division,Score))
Age9m = subset(m2009, , select=c(project,school,division,Score))
rm(m2006, m2007, m2008, m2009)

# AoE data - Get average & stdev of scores for each Judge & inAoE combination
m2006 <- aggregate(AoE2006, by=list(judge=AoE2006$Judge, IN=AoE2006$inAoE), mean)
s2006 <- aggregate(AoE2006, by=list(judge=AoE2006$Judge, IN=AoE2006$inAoE), sd)
m2007 <- aggregate(AoE2007, by=list(judge=AoE2007$Judge, IN=AoE2007$inAoE), mean)
s2007 <- aggregate(AoE2007, by=list(judge=AoE2007$Judge, IN=AoE2007$inAoE), sd)
m2008 <- aggregate(AoE2008, by=list(judge=AoE2008$Judge, IN=AoE2008$inAoE), mean)
s2008 <- aggregate(AoE2008, by=list(judge=AoE2008$Judge, IN=AoE2008$inAoE), sd)
m2009 <- aggregate(AoE2009, by=list(judge=AoE2009$Judge, IN=AoE2009$inAoE), mean)
s2009 <- aggregate(AoE2009, by=list(judge=AoE2009$Judge, IN=AoE2009$inAoE), sd)

# AoE data - Trim unneeded columns from aggregated dataframes & cleanup objects
AoE6m = subset(m2006, , select=c(judge,IN,score))
```

```
AoE6s = subset(s2006, , select=c(judge,IN,score))
AoE7m = subset(m2007, , select=c(judge,IN,score))
AoE7s = subset(s2007, , select=c(judge,IN,score))
AoE8m = subset(m2008, , select=c(judge,IN,score))
AoE8s = subset(s2008, , select=c(judge,IN,score))
AoE9m = subset(m2009, , select=c(judge,IN,score))
AoE9s = subset(s2009, , select=c(judge,IN,score))
rm(m2006, s2006, m2007, s2007, m2008, s2008, m2009, s2009)

# Edu data - Get average & stdev of scores for each judge
m2006 <- aggregate(Edu2006, by=list(judge=Edu2006$Judge, degree=Edu2006$Degree), mean)
s2006 <- aggregate(Edu2006, by=list(judge=Edu2006$Judge, degree=Edu2006$Degree), sd)
m2007 <- aggregate(Edu2007, by=list(judge=Edu2007$Judge, degree=Edu2007$Degree), mean)
s2007 <- aggregate(Edu2007, by=list(judge=Edu2007$Judge, degree=Edu2007$Degree), sd)
m2008 <- aggregate(Edu2008, by=list(judge=Edu2008$Judge, degree=Edu2008$Degree), mean)
s2008 <- aggregate(Edu2008, by=list(judge=Edu2008$Judge, degree=Edu2008$Degree), sd)
m2009 <- aggregate(Edu2009, by=list(judge=Edu2009$Judge, degree=Edu2009$Degree), mean)
s2009 <- aggregate(Edu2009, by=list(judge=Edu2009$Judge, degree=Edu2009$Degree), sd)

# Edu data - Trim unneeded columns from aggregated dataframes & cleanup objects
Edu6m = subset(m2006, , select=c(judge,degree,Score))
Edu6s = subset(s2006, , select=c(judge,degree,Score))
Edu7m = subset(m2007, , select=c(judge,degree,Score))
Edu7s = subset(s2007, , select=c(judge,degree,Score))
Edu8m = subset(m2008, , select=c(judge,degree,Score))
Edu8s = subset(s2008, , select=c(judge,degree,Score))
Edu9m = subset(m2009, , select=c(judge,degree,Score))
Edu9s = subset(s2009, , select=c(judge,degree,Score))
rm(m2006, s2006, m2007, s2007, m2008, s2008, m2009, s2009)

# I can remove these original dataframes now too
rm(Age2006,Age2007,Age2008,Age2009)
rm(AoE2006,AoE2007,AoE2008,AoE2009)
rm(Edu2006,Edu2007,Edu2008,Edu2009)
```

Descriptive Statistics

```
## Age descriptives
aggregate(Age6m, by=list(School=Age6m$school), mean)
aggregate(Age6m, by=list(School=Age6m$school), sd)
aggregate(Age7m, by=list(School=Age7m$school), mean)
aggregate(Age7m, by=list(School=Age7m$school), sd)
aggregate(Age8m, by=list(School=Age8m$school), mean)
aggregate(Age8m, by=list(School=Age8m$school), sd)
aggregate(Age9m, by=list(School=Age9m$school), mean)
aggregate(Age9m, by=list(School=Age9m$school), sd)

## AoE descriptives
aggregate(AoE6m, by=list(within=AoE6m$IN), mean)
aggregate(AoE6m, by=list(within=AoE6m$IN), sd)
aggregate(AoE7m, by=list(within=AoE7m$IN), mean)
aggregate(AoE7m, by=list(within=AoE7m$IN), sd)
aggregate(AoE8m, by=list(within=AoE8m$IN), mean)
aggregate(AoE8m, by=list(within=AoE8m$IN), sd)
aggregate(AoE9m, by=list(within=AoE9m$IN), mean)
aggregate(AoE9m, by=list(within=AoE9m$IN), sd)
aggregate(AoE6s, by=list(within=AoE6s$IN), mean)
aggregate(AoE6s, by=list(within=AoE6s$IN), sd)
aggregate(AoE7s, by=list(within=AoE7s$IN), mean)
aggregate(AoE7s, by=list(within=AoE7s$IN), sd)
aggregate(AoE8s, by=list(within=AoE8s$IN), mean)
aggregate(AoE8s, by=list(within=AoE8s$IN), sd)
aggregate(AoE9s, by=list(within=AoE9s$IN), mean)
```

```
aggregate(AoE9s, by=list(within=AoE9s$IN), sd)

## Edu descriptives
aggregate(Edu6m, by=list(Degree=Edu6m$degree), mean)
aggregate(Edu6m, by=list(Degree=Edu6m$degree), sd)
aggregate(Edu7m, by=list(Degree=Edu7m$degree), mean)
aggregate(Edu7m, by=list(Degree=Edu7m$degree), sd)
aggregate(Edu8m, by=list(Degree=Edu8m$degree), mean)
aggregate(Edu8m, by=list(Degree=Edu8m$degree), sd)
aggregate(Edu9m, by=list(Degree=Edu9m$degree), mean)
aggregate(Edu9m, by=list(Degree=Edu9m$degree), sd)
aggregate(Edu6s, by=list(Degree=Edu6s$degree), mean)
aggregate(Edu6s, by=list(Degree=Edu6s$degree), sd)
aggregate(Edu7s, by=list(Degree=Edu7s$degree), mean)
aggregate(Edu7s, by=list(Degree=Edu7s$degree), sd)
aggregate(Edu8s, by=list(Degree=Edu8s$degree), mean)
aggregate(Edu8s, by=list(Degree=Edu8s$degree), sd)
aggregate(Edu9s, by=list(Degree=Edu9s$degree), mean)
aggregate(Edu9s, by=list(Degree=Edu9s$degree), sd)
```

Area of Expertise Analysis

```
## Tests for normality (Shapiro-Wilk) & homogeneity of variance (Fligner-Killeen)
shapiro.test(as.vector( subset(AoE6m,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE6m,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE6m)
shapiro.test(as.vector( subset(AoE7m,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE7m,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE7m)
shapiro.test(as.vector( subset(AoE8m,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE8m,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE8m)
shapiro.test(as.vector( subset(AoE9m,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE9m,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE9m)
shapiro.test(as.vector( subset(AoE6s,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE6s,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE6s)
shapiro.test(as.vector( subset(AoE7s,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE7s,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE7s)
shapiro.test(as.vector( subset(AoE8s,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE8s,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE8s)
shapiro.test(as.vector( subset(AoE9s,IN==0,select=c(score))$score ))
shapiro.test(as.vector( subset(AoE9s,IN==1,select=c(score))$score ))
fligner.test(score~IN, data=AoE9s)
##          ok for t-test:  AoE6m AoE7m AoE7s
##    Wilcoxon Signed Rank:  AoE6s AoE8m AoE8s AoE9m AoE9s

## t-tests
t.test(subset(AoE6m,IN==0,select=c(score))$score,subset(AoE6m,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE7m,IN==0,select=c(score))$score,subset(AoE7m,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE7s,IN==0,select=c(score))$score,subset(AoE7s,IN==1,select=c(score))$score,paired=1)

## Wilcoxon Signed Rank (Matched Pairs) tests
wilcox.test(subset(AoE6s,IN==0,select=c(score))$score,subset(AoE6s,IN==1,select=c(score))
$score,paired=1)
wilcox.test(subset(AoE8m,IN==0,select=c(score))$score,subset(AoE8m,IN==1,select=c(score))
$score,paired=1)
wilcox.test(subset(AoE8s,IN==0,select=c(score))$score,subset(AoE8s,IN==1,select=c(score))
$score,paired=1)
wilcox.test(subset(AoE9m,IN==0,select=c(score))$score,subset(AoE9m,IN==1,select=c(score))
```

```
$score,paired=1)
wilcox.test(subset(AoE9s,IN==0,select=c(score))$score,subset(AoE9s,IN==1,select=c(score))
$score,paired=1)

## Additional t-tests
t.test(subset(AoE6s,IN==0,select=c(score))$score,subset(AoE6s,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE8m,IN==0,select=c(score))$score,subset(AoE8m,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE8s,IN==0,select=c(score))$score,subset(AoE8s,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE9m,IN==0,select=c(score))$score,subset(AoE9m,IN==1,select=c(score))$score,paired=1)
t.test(subset(AoE9s,IN==0,select=c(score))$score,subset(AoE9s,IN==1,select=c(score))$score,paired=1)
```

Education Level Analysis

```
## Tests for normality (Shapiro-Wilk) & homogeneity of variance (Fligner-Killeen)
shapiro.test(as.vector( subset(Edu6m,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu6m,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu6m,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu6m)
shapiro.test(as.vector( subset(Edu7m,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu7m,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu7m,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu7m)
shapiro.test(as.vector( subset(Edu8m,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu8m,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu8m,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu8m)
shapiro.test(as.vector( subset(Edu9m,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu9m,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu9m,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu9m)
shapiro.test(as.vector( subset(Edu6s,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu6s,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu6s,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu6s)
shapiro.test(as.vector( subset(Edu7s,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu7s,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu7s,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu7s)
shapiro.test(as.vector( subset(Edu8s,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu8s,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu8s,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu8s)
shapiro.test(as.vector( subset(Edu9s,degree=="B",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu9s,degree=="MS",select=c(Score))$Score ))
shapiro.test(as.vector( subset(Edu9s,degree=="PhD",select=c(Score))$Score ))
fligner.test(Score~degree, data=Edu9s)
##             ok for ANOVA:  Edu6m Edu7m Edu7s Edu8s Edu9m
##  need to use Kruskal-Wallis:  Edu6s Edu8m Edu9s

## ANOVAs
summary(aov(Score ~ degree, data=Edu6m))
summary(aov(Score ~ degree, data=Edu7m))
summary(aov(Score ~ degree, data=Edu7s))
summary(aov(Score ~ degree, data=Edu8s))
summary(aov(Score ~ degree, data=Edu9m))

## Kruskal-Wallis tests
kruskal.test(Edu6s$Score~Edu6s$degree)
kruskal.test(Edu8m$Score~Edu8m$degree)
kruskal.test(Edu9s$Score~Edu9s$degree)

## Additional ANOVAs
summary(aov(Score ~ degree, data=Edu6s))
```

```
summary(aov(Score ~ degree, data=Edu8m))
summary(aov(Score ~ degree, data=Edu9s))
```

Age Analysis

```
library(lme4)
library(languageR)
library(RLRsim)

# Fit using REML, then full ML, then without fixed effects (division)
A6.reml <- lmer(Score ~ division + (1|school), data=Age6m, REML=TRUE)
A7.reml <- lmer(Score ~ division + (1|school), data=Age7m, REML=TRUE)
A8.reml <- lmer(Score ~ division + (1|school), data=Age8m, REML=TRUE)
A9.reml <- lmer(Score ~ division + (1|school), data=Age9m, REML=TRUE)
A6.full <- lmer(Score ~ division + (1|school), data=Age6m, REML=FALSE)
A7.full <- lmer(Score ~ division + (1|school), data=Age7m, REML=FALSE)
A8.full <- lmer(Score ~ division + (1|school), data=Age8m, REML=FALSE)
A9.full <- lmer(Score ~ division + (1|school), data=Age9m, REML=FALSE)
A6.noDiv <- lmer(Score ~ 1|school, data=Age6m, REML=FALSE)          # no fixed (division)
A7.noDiv <- lmer(Score ~ 1|school, data=Age7m, REML=FALSE)          # no fixed (division)
A8.noDiv <- lmer(Score ~ 1|school, data=Age8m, REML=FALSE)          # no fixed (division)
A9.noDiv <- lmer(Score ~ 1|school, data=Age9m, REML=FALSE)          # no fixed (division)

# Plot random effects (school contributions)
dotplot(ranef(A6.reml, postVar=TRUE), scales=list(cex=.6))
dotplot(ranef(A7.reml, postVar=TRUE), scales=list(cex=.6))
dotplot(ranef(A8.reml, postVar=TRUE), scales=list(cex=.6))
dotplot(ranef(A9.reml, postVar=TRUE), scales=list(cex=.6))

# LR tests for random effects (schools)
exactRLRT(A6.reml, nsim=100000)
exactRLRT(A7.reml, nsim=100000)
exactRLRT(A8.reml, nsim=100000)
exactRLRT(A9.reml, nsim=100000)

# Display empirical fixed effect (division) means
tapply(Age6m$Score, Age6m$division, mean)
tapply(Age7m$Score, Age7m$division, mean)
tapply(Age8m$Score, Age8m$division, mean)
tapply(Age9m$Score, Age9m$division, mean)

# ANOVA tests for fixed effects (division)
anova(A6.full, A6.noDiv)
anova(A7.full, A7.noDiv)
anova(A8.full, A8.noDiv)
anova(A9.full, A9.noDiv)

# Bayesian 95% CIs for fixed effects (division)
A6post <- mcmcsamp(A6.reml, 100000, withMCMC=TRUE)
A7post <- mcmcsamp(A7.reml, 100000, withMCMC=TRUE)
A8post <- mcmcsamp(A8.reml, 100000, withMCMC=TRUE)
A9post <- mcmcsamp(A9.reml, 100000, withMCMC=TRUE)
HPDinterval(A6post)
HPDinterval(A7post)
HPDinterval(A8post)
HPDinterval(A9post)
```

Age Analysis – Bayesian Confidence Intervals

```
# Calculate midpoints & standard errors for previously rendered Bayesian CIs
```

```
B7E.m = (84.398 + 78.887)/2
B7E.s = (84.398 - 78.887)/(2 * 1.96)
B7J.m = (82.7468 + 69.701)/2
B7J.s = (82.7468 - 69.701)/(2 * 1.96)
B7S.m = (84.2668 + 70.047)/2
B7S.s = (84.2668 - 70.047)/(2 * 1.96)
B8E.m = (82.5266 + 77.782)/2
B8E.s = (82.5266 - 77.782)/(2 * 1.96)
B8J.m = (81.2561 + 69.68)/2
B8J.s = (81.2561 - 69.68)/(2 * 1.96)
B8S.m = (82.1106 + 69.589)/2
B8S.s = (82.1106 - 69.589)/(2 * 1.96)

# Test to see if 2007 Bayesian CI midpoints are significantly different at alpha=.05
 # Elementary vs Junior
 (2*sqrt(B7E.s^2 + B7J.s^2) < B7J.m - B7E.m) && (B7J.m - B7E.m < 2*B7J.s + 2*B7E.s)
 # Elementary vs Senior
 (2*sqrt(B7E.s^2 + B7S.s^2) < B7S.m - B7E.m) && (B7S.m - B7E.m < 2*B7S.s + 2*B7E.s)
 # Junior vs Senior
 (2*sqrt(B7J.s^2 + B7S.s^2) < B7S.m - B7J.m) && (B7S.m - B7J.m < 2*B7S.s + 2*B7J.s)

# Test to see if 2008 Bayesian CI midpoints are significantly different at alpha=.05
 # Elementary vs Junior
 (2*sqrt(B8E.s^2 + B8J.s^2) < B8J.m - B8E.m) && (B8J.m - B8E.m < 2*B8J.s + 2*B8E.s)
 # Elementary vs Senior
 (2*sqrt(B8E.s^2 + B8S.s^2) < B8S.m - B8E.m) && (B8S.m - B8E.m < 2*B8S.s + 2*B8E.s)
 # Junior vs Senior
 (2*sqrt(B8J.s^2 + B8S.s^2) < B8S.m - B8J.m) && (B8S.m - B8J.m < 2*B8S.s + 2*B8J.s)
```

Judge Deviation Analysis

```
Dev6 <- Dev2006[order(-Dev2006$AveDiff),]
Dev7 <- Dev2007[order(-Dev2007$AveDiff),]
Dev8 <- Dev2008[order(-Dev2008$AveDiff),]
Dev9 <- Dev2009[order(-Dev2009$AveDiff),]
rm(Dev2006,Dev2007,Dev2008,Dev2009)
```

Fun Facts from Science Fair Projects

Corn grows better upside-down.

To get the most popped kernels, store your popcorn in the freezer.

Cheaper BBs seem to hit their targets more accurately than expensive BBs.

DVR technology like “TiVo” can be applied to a radio, which can then record and play back songs, skip commercials, etc.

Fast food not only makes rats lethargic, but smell worse too.

Soap and water work better than hand sanitizers.

Air fresheners facilitate the growth of bacteria in water.

There are junior high students who can decode the RSA encryption used by banks.
(Maybe that one isn't so fun.)

References

Baayen, R. H. (2009). languageR (Version 0.955) [Software]. Available from http://cran.r-project.org/web/packages/languageR/index.html

Bates, Douglas M. (2006, Oct. 20). "[R] mcmcsamp - How does it work?" [Msg 2]. Message posted to http://n4.nabble.com/R-mcmcsamp-How-does-it-work-td810547.html#a810547

Bates, Douglas M., & Maechler, Martin (2010). lme4 (Version 0.999375-33) [Software]. Available from http://cran.r-project.org/web/packages/lme4/index.html

Cox, Jimmy (2007 October 31). http://www.articlexplosion.com/articledetail.php?artid=31301&catid=159&title=A+History+of+Science+Fairs

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Harcourt Brace Jovanovich.

Dallal, Gerard E. (2007). The Little Handbook of Statistical Practice. Retrieved from March, 2010, from http://www.tufts.edu/~gdallal/LHSP.HTM

E. W. Scripps. (2009). In Wikipedia, the free encyclopedia. Retrieved February, 2009, from http://en.wikipedia.org/wiki/E._W._Scripps

Edward Willis Scripps. (2009). In Encyclopædia Britannica. Retrieved March 27, 2009, from Encyclopædia Britannica Online: http://www.britannica.com/EBchecked/topic/529991/Edward-Willis-Scripps

Fredrickson, Clifford T., & Mikkelson, Mary Domb (1979). The Science Fair: Making It Work, Making It Fair. *The American Biology Teacher*, 41(8), 449-550+504-505. http://www.jstor.org/stable/4446728

McBurney, Wendell F. (1978). The Science Fair: A Critique and Some Suggestions. *The American Biology Teacher*, 40(7), 419-422. http://www.jstor.org/stable/4446323

Pinheiro, José C., & Bates, Douglas M. (2002). Mixed-Effects Models in S and S-PLUS. New York: Springer-Verlag.

Scheipl, Fabian. (2010). RLRsim (Version 2.0-4) [Software]. Available from http://cran.r-project.org/web/packages/RLRsim/index.html

Scheipl, Fabian. (2010). Exact (Restricted) Likelihood Ratio tests for mixed and additive models. CRAN Repository. Retrieved February 12, 2010, from http://cran.r-project.org/web/packages/RLRsim/RLRsim.pdf.

Science Fair. (2009). In Wikipedia, the free encyclopedia. Retrieved January, 2009, from http://en.wikipedia.org/wiki/Science_fair

"Science Fair History" (2008, July-August). Copyright 2003-2009 Super Science Fair Projects – All Rights Reserved. http://www.super-science-fair-projects.com/science-fair-history.html

Smith, Carol L., Maclin, Deborah, Houghton, Carolyn, & Hennessey, M. Gertrude. (2000). Sixth-Grade Students' Epistemologies of Science: The Impact of School Science Experiences on Epistemological Development. *Cognition and Instruction*, 18 (3), 349-422.

Society for Science & the Public. (2009). History of Society for Science & the Public. Retrieved February, 2009, from http://sciserv.org/history.html

Twisk, Jos W. R. (2006) *Applied Multilevel Analysis*, Cambridge, UK : Cambridge University Press

William Emerson Ritter. (2009). In Wikipedia, the free encyclopedia. Retrieved February, 2009, from http://en.wikipedia.org/wiki/William_Emerson_Ritter