Here's head judge Tom Colicchio explaining what helps eliminate bias when judges evaluate the chef testants' food:

Here's what does [eliminate propensities for bias]: judging the food on particular criteria. And here are the criteria we use: First and foremost, when tasting the food we look to see if, technically, it was prepared correctly or whether it was overcooked or undercooked. After that, we check to see whether it was correctly seasoned, by which I'm talking about whether it was salted correctly, because salt has the ability to bring out the other three types of taste you experience on your tongue, i.e., sweetness, bitterness and sourness. Then we look at how items are cut. Are they cut evenly? If so, they will cook evenly. We look at food combinations to see if the proportions are harmonious. And lastly, we look at presentation, but usually only when it is particularly ugly. If veggies are cooked correctly, they'll stay green; if not, they'll turn brown. How something is cut will affect presentation. We also just take note of whether, as with all great chefs, a personal style is emerging in a consistent way, or whether they're just all over the place. Often we've seen a chef come in with a particular style and then, part-way through the competition, begin mimicking everyone else. These chefs tend to flame out; they don't make it to the final four, and, frankly, they're not yet secure and mature enough as chefs to be there. We do look at originality, as with Bryan’s winning take on chips and guacamole in Episode Two, or Kevin’s bacon jam, which was utterly original, different, and very, very good. I knew exactly where Bryan’s dish for Joel Robuchon came from – he adapted a dish from Thomas Keller – but he did make it his own. And, even hearkening back to prior seasons, most of our viewers were not familiar with molecular gastronomy and thought that Marcel was innovating, whereas, in fact, his techniques had been around for at least a decade and he wasn’t being particular novel in his application of it but was solidly adept at what he was doing.

You'll notice that we judges are seldom in disagreement. This is because we are always applying the criteria I just outlined above, and, in doing so, tend to reach similar conclusions. We're not applying whim or personal preference; the dishes themselves tend to give each of us the same basic information upon which to base our decisions.

Using the information provided above, let's map out a rough rubric for Top Chef judging on the following page.

- **What are the main criteria or traits that should be used for evaluating each dish?**
- **Can you describe the different levels of performance on each trait based on the information Chef Colicchio provides?**
<table>
<thead>
<tr>
<th>Criteria</th>
<th>Exemplary</th>
<th>Accomplished</th>
<th>Emerging</th>
<th>Beginning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deconstructing Rubrics

**Definition:** A scoring tool that lays out the specific expectations for an activity or product. A rubric is an authentic assessment tool used to measure student's work.

Why Use A Rubric?

### Task Description

You will write a reflective essay. A reflective essay is a piece of writing that basically involves your views and feelings about a particular subject. The goal of a reflective essay is to not only discuss what you learned, but to convey the personal experiences and findings that resulted.

### Scale Levels

### Dimensions

<table>
<thead>
<tr>
<th>Reflective Essay</th>
<th>Name:</th>
<th>Date:</th>
<th>Score:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reflect personal learning stretch in Science Project</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shows great depth of knowledge and learning, reveals feelings and thoughts, abstract ideas reflected through use of specific details.</td>
<td>Exceeds Standard</td>
<td>Meets Standard</td>
<td>Nearly Meets Standard</td>
</tr>
<tr>
<td>Relates learning with research and project, personal and general reflections included, uses concrete language.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does not go deeply into the reflection of learning, generalizations and limited insight, uses some detail.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little or no explanation or reflection on learning, no or few details to support reflection.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shows no evidence of learning or reflection.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Organization-Structural Development of the Idea</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writer demonstrates logical and subtle sequencing of ideas through well-developed paragraphs; transitions are used to enhance organization.</td>
<td>Exceeds Standard</td>
<td>Meets Standard</td>
<td>Nearly Meets Standard</td>
</tr>
<tr>
<td>Paragraph development present but not perfected.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical organization; organization of ideas not fully developed.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No evidence of structure or organization.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conclusion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The conclusion is engaging and restates personal learning.</td>
<td>Exceeds Standard</td>
<td>Meets Standard</td>
<td>Nearly Meets Standard</td>
</tr>
<tr>
<td>The conclusion restates the learning.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The conclusion does not adequately restate the learning.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incomplete and/or unfocused.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mechanics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No errors in punctuation, capitalization and spelling.</td>
<td>Exceeds Standard</td>
<td>Meets Standard</td>
<td>Nearly Meets Standard</td>
</tr>
<tr>
<td>Almost no errors in punctuation, capitalization and spelling.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many errors in punctuation, capitalization and spelling.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numerous and distracting errors in punctuation, capitalization and spelling.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not applicable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Usage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No errors in sentence structure and word usage.</td>
<td>Exceeds Standard</td>
<td>Meets Standard</td>
<td>Nearly Meets Standard</td>
</tr>
<tr>
<td>Almost no errors in sentence structure and word usage.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many errors in sentence structure and word usage.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numerous and distracting errors in sentence structure and word usage.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not applicable</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Grading – consistency, ease, efficiency
• Communicate Expectations to Students
• Student Feedback Tool
• Assessment Tool for Documenting Learning – individual students
• Assessment Tool for Documenting Learning – groups of students (aggregated data)
• Granularity – dimensions (competencies) and scales (levels of learning)
• Reflective Tool for Students
• Reflective Tool for Teachers
Levels of Performance

Frequency: 5-Point Scale
1 – Never
2 – Rarely
3 – Occasionally
4 – Often
5 – Always

Level of Satisfaction: 4-Point Scale
1 – Very Dissatisfied
2 – Dissatisfied
3 – Satisfied
4 – Very Satisfied

Level of Satisfaction: 5-Point Scale
1 – Very Dissatisfied
2 – Dissatisfied
3 – Satisfied
4 – Neutral/Unsure
5 – Very Satisfied

Level of Quality: 4-Point Scale
1 – Poor
2 – Fair
3 – Good
4 – Excellent

Level of Familiarity: 5-Point Scale
1 – Not at all Familiar
2 – Slightly Familiar
3 – Somewhat Familiar
4 – Moderately Familiar
5 – Extremely Familiar

Frequency: 7-Point Scale
1 – Never
2 – Rarely (approx. 10%)
3 – Occasionally (approx. 30%)
4 – Sometimes (approx. 50%)
5 – Frequently (approx. 70%)
6 – Usually (approx. 90%)
7 – Every time

Level of Satisfaction: 7-Point Scale
1 – Completely Dissatisfied
2 – Mostly Dissatisfied
3 – Somewhat Dissatisfied
4 – Neither Satisfied or Dissatisfied
5 – Somewhat Satisfied
6 – Mostly Satisfied
7 – Completely Satisfied

Level of Awareness: 5-Point Scale
1 – Not at all Aware
2 – Slightly Aware
3 – Somewhat Aware
4 – Moderately Aware
5 – Extremely Aware

Level of Skill: 5-Point Scale
1 – No Skill
2 – Low Skilled
3 – Neutral
4 – Skilled
5 – Very Skilled

Response Anchors
Types of Rubrics

Analytic: categorizes and scores components of the activity/product.
- when there are many dimensions to consider OR when dimensions are weighted differently

Holistic: scores the activity/product as a whole.
- quick judgments OR evaluated performance criteria cannot be easily separated

Holistic Rubric Example - Oral Report

5 - Excellent
The student clearly describes the question studied and provides strong reasons for its importance. Specific information is given to support the conclusions that are drawn and described. The delivery is engaging and sentence structure is consistently correct. Eye contact is made and sustained throughout the presentation. There is strong evidence of preparation, organization, and enthusiasm for the topic. The visual aid is used to make the presentation more effective. Questions from the audience are clearly answered with specific and appropriate information.

4 - Very Good
The student described the question studied and provides reasons for its importance. An adequate amount of information is given to support the conclusions that are drawn and described. The delivery and sentence structure are generally correct. There is evidence of preparation, organization, and enthusiasm for the topic. The visual aid is mentioned and used. Questions from the audience are answered clearly.

3 - Good
The student describes the question studied and conclusions are stated, but supporting information is not as strong as a 4 or 5. The delivery and sentence structure are generally correct. There is some indication of preparation and organization. The visual aid is mentioned. Questions from the audience are answered.

2 - Limited
The student states the question studied, but fails to fully describe it. No conclusions are given to answer the question. The delivery and sentence structure is understandable, but with some errors. Evidence of preparation and organization is lacking. The visual aid may or may not be mentioned. Questions from the audience are answered with only the most basic response.

1 - Poor
The student makes a presentation without stating the question or its importance. The topic is unclear and no adequate conclusions are stated. The delivery is difficult to follow. There is no indication of preparation or organization. Questions from the audience receive only the most basic, or no, response.

0 - No oral presentation is attempted.

Resource: http://www.middleweb.com/rubricsHG.html
Constructing Rubrics

Constructing Rubrics: Four-Step Model
- Step 1 – Reflecting
- Step 2 – Defining Levels of Performance
- Step 3 – Grouping Criteria/Defining Dimensions
- Step 4 – Testing the Rubric

Step 1 – Reflecting
Take time to reflect on what you want from the students, why you created this activity, and what your expectations are:

- Why did you create this activity?
- Have you given this assignment or a similar assignment before?
- Do the students already possess the skills needed to complete the activity?
- What exactly is the task assigned?
- What would you consider evidence that the students will provide to show that they have accomplished what you hoped they would accomplish (i.e., the outcomes)?
- What are the highest expectations you have for student performance on this activity?
- What is the worst fulfillment of the assignment, short of simply not turning it in at all?

Step 2 – Define the Levels of Performance
- Add a description of the highest level of performance for each outcome listed.
- Add a description of the lowest level of performance for each outcome listed.
- Add descriptions for the intermediate performance levels.

Step 3 – Grouping Criteria/Defining Dimensions
- Group similar expectations together and label each set of grouping.
- Transfer your lists and groupings to a rubric grid.
- Labels for the groups of performance expectations now become the dimensions of the rubric.

Step 4 – Testing the Rubric
- Apply the rubric to actuals examples of student work. If possible, apply to a wide range of student work.
- Share and discuss with colleagues, revise as appropriate.
- Hold a norming session to ensure faculty members are interpreting the rubric in the same way. This process calibrates the use of the rubric making it a more reliable and valid assessment tool.

Source:
**Guidelines Norming Session**  
Adapted from University of Hawaii, Manoa (http://manoa.hawaii.edu/assessment/howto/rubrics.htm)

Materials & Resource Needed:  
Copies of the rubric and score sheets  
Examples of poor, average and good student work to assess with rubric

Process:  
1. Describe the purpose of the activity, stressing how it fits into program assessment plans. Explain that the purpose is to assess the program, not individual students or faculty, and describe ethical guidelines, including respect for confidentiality and privacy.

2. Describe the nature of the products that will be reviewed, briefly summarizing how they were obtained.

3. Describe the scoring rubric and its categories. Explain how it was developed.  
   - Analytic: Explain that readers should rate each dimension of an analytic rubric separately, and they should apply the criteria without concern for how often each score (level of mastery) is used.  
   - Holistic: Explain that readers should assign the score or level of mastery that best describes the whole piece; some aspects of the piece may not appear in that score and that is okay. They should apply the criteria without concern for how often each score is used.

4. Give each scorer a copy of several student products that are exemplars of different levels of performance. Ask each scorer to independently apply the rubric to each of these products, writing their ratings on a scrap sheet of paper.

5. Once everyone is done, collect everyone's ratings and display them so everyone can see the degree of agreement. This is often done on a blackboard, with each person in turn announcing his/her ratings as they are entered on the board. Alternatively, the facilitator could ask raters to raise their hands when their rating category is announced, making the extent of agreement very clear to everyone and making it very easy to identify raters who routinely give unusually high or low ratings.

6. Guide the group in a discussion of their ratings. There will be differences. This discussion is important to establish standards. Attempt to reach consensus on the most appropriate rating for each of the products being examined by inviting people who gave different ratings to explain their judgments. Raters should be encouraged to explain by making explicit references to the rubric and the student work. Usually consensus is possible, but sometimes a split decision is developed, e.g., the group may agree that a product is a "3-4" split because it has elements of both categories. This is usually not a problem. You might allow the group to revise the rubric to clarify its use but avoid allowing the group to drift away from the rubric and learning outcome(s) being assessed.

7. Once the group is comfortable with how the rubric is applied, the rating begins. Explain how to record ratings using the score sheet and explain the procedures. Reviewers begin scoring.
8. If you can quickly summarize the scores, present a summary to the group at the end of the reading. You might end the meeting with a discussion of five questions:
   - Are results sufficiently reliable?
   - What do the results mean? Are we satisfied with the extent of students' learning?
   - Who needs to know the results?
   - What are the implications of the results for curriculum, pedagogy, or student support services?
   - How might the assessment process, itself, be improved?

Example Data from Norming Session
   - Rubric with a 5-point scale
   - Normed three student essays that reflect good, average, poor student work

<table>
<thead>
<tr>
<th></th>
<th>Student Essay 1</th>
<th>Student Essay 2</th>
<th>Student Essay 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer 1</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 3</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 4</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 5</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Reviewer 6</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 7</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 8</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Reviewer 9</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer 10</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Totals</th>
<th>7-4's</th>
<th>3-5's</th>
<th>7-3's</th>
<th>2-4's</th>
<th>1-5</th>
<th>8-1's</th>
<th>1-2</th>
<th>1-3</th>
</tr>
</thead>
</table>

Possible statements:

- Reviewers were within one-point of agreement 93% of the time
- Reviewers agreed 73% of the time

...Other considerations?
**Rubric for Assessing Chocolate**

<table>
<thead>
<tr>
<th></th>
<th>5 Excellent</th>
<th>4 Good</th>
<th>3 Good</th>
<th>2 Unacceptable</th>
<th>Product #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Appearance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Packaging</td>
<td>Aesthetically stylistic and appealing; simple to open and access.</td>
<td>Nice, simple, clean appearance.</td>
<td>Packaging protects the product, but is spartan and minimally appealing.</td>
<td>Product is not secure, has potential to come out of package.</td>
<td></td>
</tr>
<tr>
<td><strong>Color, Appearance</strong></td>
<td>Product appearance exudes quality, openly invites one to consume.</td>
<td>Product is appealing. Looks like how chocolate should taste.</td>
<td>Product resembles chocolate in some ways, but color or shape are dull and pedestrian.</td>
<td>Product does not look like chocolate or look like it will taste like chocolate.</td>
<td></td>
</tr>
<tr>
<td><strong>Flavor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texture, Feel (in mouth)</td>
<td>Soft, supple feel in mouth, velvet smoothness.</td>
<td>Texture and feel are solidly chocolate.</td>
<td>Feels chocolate-like, but not quite all the way.</td>
<td>Doesn’t feel at all like chocolate on palate; could be anything but chocolate.</td>
<td></td>
</tr>
<tr>
<td>Taste</td>
<td>Full, rich chocolate flavor; lasting, round, deep finish.</td>
<td>Solid chocolate flavor, pleasant and agreeable.</td>
<td>Chocolate taste, for sure, but not pronounced or deep.</td>
<td>Does not taste at all like chocolate.</td>
<td></td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>The product is extremely satisfying.</td>
<td>The product is satisfying.</td>
<td>The product is not completely satisfying.</td>
<td>The product is not satisfying at all.</td>
<td></td>
</tr>
</tbody>
</table>
Abstract  Reviewers of research reports frequently criticize the choice of statistical methods. While some of these criticisms are well-founded, frequently the use of various parametric methods such as analysis of variance, regression, correlation are faulted because: (a) the sample size is too small, (b) the data may not be normally distributed, or (c) The data are from Likert scales, which are ordinal, so parametric statistics cannot be used. In this paper, I dissect these arguments, and show that many studies, dating back to the 1930s consistently show that parametric statistics are robust with respect to violations of these assumptions. Hence, challenges like those above are unfounded, and parametric methods can be utilized without concern for “getting the wrong answer”.

Keywords  Likert · Statistics · Robustness · ANOVA

One recurrent frustration in conducting research in health sciences is dealing with the reviewer who decides to take issue with the statistical methods employed. Researchers do occasionally commit egregious errors, usually the multiple test phenomenon associated with data—dredging. But this is rarely the basis of reviewer’s challenges. As Bacchetti (2002) has pointed out, many of these comments are unfounded or wrong, and appear to result from a review culture that encourages “overvaluation of criticism for its own sake, inappropriate statistical dogmatism”, and is subject to “time pressure, and lack of rewards for good peer reviewing”. Typical reviewers’ comments in this genre may resemble those listed below, drawn from reviews of 5 different papers, all brought to my attention in a 2 month period:

Paper 1

…and in case of use of parametric tests (like t-test) I’d like to see the results of the assumption of normality of the distribution

G. Norman
McMaster University, 1200 Main St. W., Hamilton, ON L8N3Z5, Canada
e-mail: norman@mcmaster.ca

Published online: 10 February 2010
Paper 2

...the authors [use] analytical practices which are not supported by the type of data they have available.... Ordinal data do not support mathematical calculations such as change scores, .... the approach adopted by the authors is indefensible....

Paper 3

The statistical analysis of correlation .... is done with a method not suitable for non-parametric, consult with statistician. The t-test performed requires that the data be normally distributed. However, the validity of these assumptions ...has not been justified. Given the small number of participants in each group, can the authors claim statistical significance?

Paper 4:

The sample size is very low .... As the data was not drawn from a normal distribution due to the very low sample size, it is not possible to analyse the data using parametric tests, such as ANOVA.

Paper 5:

Did you complete a power analysis to determine if your N was high enough to do these tests? ...with the low N, not sure if you can claim significance without a power analysis to confirm; otherwise Type II error is highly possible in your results

Some of these comments, like the proscription on the use of ANOVA with small samples, the suggestion to use power analysis to determine if sample size was large enough to do a parametric test, or the concern that a significant result still might be a Type II error, are simply wrong and reveal more about the reviewer’s competence than the study design.

Others, like the various distributional assumptions or the use of parametric statistics with ordinal data, may be strictly true, but fail to account for the robustness of parametric tests, and ignore a substantial literature suggesting that parametric statistics are perfectly appropriate. Regrettably, these reviewers can find compatible company in the literature. For example, Kuzon et al. (1996) writes about the “seven deadly sins of statistical analysis”. Sin 1 is using parametric statistics on ordinal data; Sin 2 relates to the assumption of normality and claims that “Before parametric statistical analysis is appropriate... the study sample must be drawn from a normally distributed population [ital. theirs]” and (2) the sample size must be large enough to be representative of the population”.1

The intention of this paper is to redress the balance. One of the beauties of statistical methods is that, although they often involve heroic assumptions about the data, it seems to matter very little even when these are violated. In order to help researchers more effectively deal with challenges like those above, this paper is a review of the assumptions of

---

1 Representativeness is required of all statistical tests and is fundamental to statistical inference. But it is unrelated to sample size.
Various statistical methods and the problems (or more commonly the lack of problems) when the assumptions are violated.

These issues are particularly germane to educational research because so many of our studies involve rating scales of one kind or another and virtually all rating scales involve variants on the 7 point Likert scale. It does not take a lot of thought to recognize that Likert scales are ordinal. To quote a recent article in Medical Education (Jamieson 2004) “the response categories have a rank order but the intervals between values cannot be presumed equal”. True—strictly speaking. The consequence is that, again according to Jamieson, “the appropriate descriptive and inferential statistics differ for ordinal and interval variables and if the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion”. Again, true—strictly speaking. But what is left unsaid is how much it increases the chance of an erroneous conclusion. This is what statisticians call “robustness”, the extent to which the test will give the right answer even when assumptions are violated. And if it doesn’t increase the chance very much (or not at all), then we can press on.

It is critically important to take this next step, not simply because we want to avoid “coming to the wrong conclusion”. As it turns out, parametric methods are incredibly versatile, powerful and comprehensive. Modern parametric statistical methods like factor analysis, hierarchical linear models, structural equation models are all based on an assumption of normally distributed, interval-level data. Similarly generalizability theory, is based on ANOVA that again is a parametric procedure. By contrast, rank methods like Spearman rho, Kruskal–Wallis, appear frozen in time and are used only rarely. They can handle only the simplest of designs. If Jamieson and others are right and we cannot use parametric methods on Likert scale data, and we have to prove that our data are exactly normally distributed, then we can effectively trash about 75% of our research on educational, health status and quality of life assessment (as pointed out by one editor in dismissing one of the reviewer comments above).

Well, despite the fact that Jamieson’s recent paper has apparently taken the medical education world by surprise and was the most downloaded paper in Medical Education in 2004, the arguments back and forth have been going on for a very long time. I will spend some time reviewing these issues, but instead of focusing on assumptions, I will directly address the issue of robustness. I will explore the impact of three characteristics-sample size, non-normality, and ordinal-level measurement, on the use of parametric methods. The arguments and responses:

1) You can’t use parametric tests in this study because the sample size is too small

This is the easiest argument to counter. The issue is not discussed in the statistics literature, and does not appear in statistics books, for one simple reason. Nowhere in the assumptions of parametric statistics is there any restriction on sample size. It is simply not true, for example, that ANOVA can only be used for large samples, and one should use a t test for smaller samples. ANOVA and t tests are based on the same assumptions; for two groups the F test from the ANOVA is the square of the t test. Nor is it the case that below some magical sample size, one should use non-parametric statistics. Nowhere is there any evidence that non-parametric tests are more appropriate than parametric tests when sample sizes get smaller.

In fact, there is one circumstance where non-parametric tests will give an answer that can be extremely conservative (i.e. wrong). The act of dichotomizing data (for example, using final exam scores to create Pass and Fail groups and analyzing failure rates, instead of simply analyzing the actual scores), can reduce statistical power enormously. Simulations I conducted showed that if the data are reasonably continuous and reasonably “well-
behaved” (begging the issue of what is “reasonable”) dichotomizing the data led to a reduction in statistical power. To do this, I began with data from two hypothetical distributions with a known separation, so that I could compute a Z test on the difference between means. (For example, two distributions centered on 50 and 55, with a sample size of 100, and a standard deviation of 15. I then drew a cutpoint so that each distribution was divided into 2 groups (a “pass and a “fail”). This then led to a 2 × 2 table with proportions derived from the overlap of the original distributions and the location of the cutpoint. I then computed the required sample size for a P-value of .05 using a standard formula. Finally, I calculated the ratio of the sample size for a significant Z test and computed the ratio. The result was a cost in sample size from 20% (when the cutpoint was on the 50th percentile) to 2,600% (when the cutpoint was at the 5th or 95th percentile). The finding is neither new nor publishable; other authors have shown similar effects (Suissa 1991; Hunter and Schmidt 1990).

Sample size is not unimportant. It may be an issue in the use of statistics for a number of reasons unrelated to the choice of test:

(a) With too small a sample, external validity is a concern. It is difficult to argue that 2 physicians or 3 nursing students are representative of anything (qualitative research notwithstanding). But this is an issue of judgment, not statistics.

(b) As we will see in the next section, when the sample size is small, there may be concern about the distributions (see next section). However, it turns out that the demarcation is about 5 per group. And the issue is not that one cannot do the test, but rather that one might begin to worry about the robustness of the test.

(c) Of course, small samples require larger effects to achieve statistical significance. But to say, as one reviewer said above, “Given the small number of participants in each group, can the authors claim statistical significance?”, simply reveals a lack of understanding. If it’s significant, it’s significant. A small sample size makes the hurdle higher, but if you’ve cleared it, you’re there.

2) You can’t use t tests and ANOVA because the data are not normally distributed

This is likely one of the most prevalent myths. We all see the pretty bell curves used to illustrate z tests, t tests and the like in statistics books, and we learn that “parametric tests are based on the assumption of normality”. Regrettably, we forget the last part of the sentence. For the standard t tests ANOVAs, and so on, it is the assumption of normality of the distribution of means, not of the data. The Central Limit Theorem shows that, for sample sizes greater than 5 or 10 per group, the means are approximately normally distributed regardless of the original distribution. Empirical studies of robustness of ANOVA date all the way back to Pearson (1931) who found ANOVA was robust for highly skewed non-normal distributions and sample sizes of 4, 5 and 10. Boneau (1960) looked at normal, rectangular and exponential distributions and sample sizes of 5 and 15, and showed that 17 of the 20 calculated P-values were between .04 and .07 for a nominal 0.05. Thus both theory and data converge on the conclusion that parametric methods examining differences between means, for sample sizes greater than 5, do not require the assumption of normality, and will yield nearly correct answers even for manifestly nonnormal and asymmetric distributions like exponentials.

3) You can’t use parametric tests like ANOVA and Pearson correlations (or regression, which amounts to the same thing) because the data are ordinal and you can’t assume normality.

The question, then, is how robust are Likert scales to departures from linear, normal distributions. There are actually three answers. The first, perhaps the least radical, is that
expounded by Carifio and Perla (2008) in their response to Jamieson (2004). They begin, as
I have, in pointing out that those who defend the logical position that parametric methods
cannot be used on ordinal data ignore the many studies of robustness. But their strongest
argument appears to be that while Likert questions or items may well be ordinal, Likert
scales, consisting of sums across many items, will be interval. It is completely analogous to
the everyday, and perfectly defensible, practice of treating the sum of correct answers on a
multiple choice test, each of which is binary, as an interval scale. The problem is that they,
by extension, support the “ordinalist” position for individual items, stating “Analyzing a
single Likert item, it should also be noted, is a practice that should occur only rarely.”
Their rejoinder can hardly be viewed as a strong refutation.

The second approach, as elaborated by Gaito (1980), is that this is not a statistics
question at all. The numbers “don’t know where they came from”. What this means is that,
even if conceptually a Likert scale is ordinal, to the extent that we cannot theoretically
guarantee that the true distance between 1 = “Definitely disagree” and 2 = “Disagree” is
the same as “4 = “No opinion” and 5 = “Moderately agree”, this is irrelevant to the
analysis because the computer has no way of affirming or denying it. There are no inde-
pendent observations to verify or refute the issue. And all the computer can do is draw
conclusions about the numbers themselves. So if the numbers are reasonably distributed,
we can make inferences about their means, differences or whatever. We cannot, strictly
speaking, make further inferences about differences in the underlying, latent, characteristic
reflected in the Likert numbers, but this does not invalidate conclusions about the numbers.
This is almost a “reductio ad absurbum” argument, and appears to solve the problem by
making it someone else’s, but not the statistician’s problem. After all, someone has to
decline whether the analysis done on the numbers reflects the underlying constructs, and
Gaito provides no support for this inference.

So let us return to the more empirical approach that has been used to investigate
robustness. As we showed earlier, ANOVA and other tests of central tendency are highly
robust to things like skewness and non-normality. Since an ordinal distribution amounts to
some kind of nonlinear relation between the number and the latent variable, then in my
view the answer to the question of robustness with respect to ordinality is essentially
answered by the studies cited above showing robustness with respect to non-normality.

However, when it comes to correlation and regression, this proscription cannot be dealt
with quite so easily. The nature of regression and correlation methods is that they inher-
ently deal with variation, not central tendency (Cronbach 1957). We are no longer talking
about a distribution of means. Rather, the magnitude of the correlation is sensitive to
individual data at the extremes of the distribution, as these “anchor” the regression line.
So, conceivably, distortions in the distribution—skewness or non-linearity—could well
“give the wrong answer”.

If the Likert ratings are ordinal which in turn means that the distributions are highly
skewed or have some other undesirable property, then it is a statistical issue about whether
or not we can go ahead and calculate correlations or regression coefficients. It again
becomes an issue of robustness. If the distributions are not normal and linear, what happens
to the correlations? This time, there is no “Central Limit Theorem” to provide theoretical
confidence. However, there have been a number of studies that are reassuring. Pearson
(1931, 1932a, b), Dunlap (1931) and Havlicek and Peterson (1976) have all shown, using
theoretical distributions, that the Pearson correlation is robust with respect to skewness and
nonnormality. Havlicek and Peterson did the most extensive simulation study, looking at
sample size from 5 to 60 (with 3,000–5,000 replications each), for normal, rectangular, and
ordinal scales (the latter obtained by adding and subtracting numbers at random). They
then computed the proportions of observed correlations within each nominal magnitude, e.g. for a nominal proportion of 0.05, the proportion of samples in this zone ranged from .046 to .053. They concluded that “The Pearson r is rather insensitive to extreme violations of the basic assumptions of normality and the type of scale”.

I confirmed these results recently with some real scale data. I had available a data set from 93 patients who had completed a quality of life measure related to cough consisting of 8, 10 point scales, on two occasions (Fletcher et al. 2010). The questions were of the form:

I have had serious health problems before my visit.
I have been unable to participate in activities before my visit.

and the responses were on a 10 point scale, with gradations:

0 = no problem
2 = mild problem
4 = moderate problem
6 = severe problem
8 = very serious problem
10 = worst possible problem

Each response was made by inspecting a card that showed: (a) The number, (b) The description, (c) A graphical “ladder”, and (d) a sad to happy face.

Using the data set, I computed the Pearson correlation between each of the Time 1 scale responses and each of the Time 2 responses, resulting in 64 correlations based on a sample of 93 respondents. I then calculated the Spearman correlation based on ranks derived from the 10 scale points. Finally, I then treated these 64 pairs of Spearman and Pearson correlations as raw data, and computed the regression line, predicting the Spearman correlation from the Pearson correlation. A perfect relationship would have a correlation (Pearson) of 1.0 between the calculated Pearson and Spearman correlations, a slope of 1.0 and an intercept of 0.0.

To then create more extremely ordinal data sets, I first turned the raw data into 5 point scales, by combining 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8, 9 and 10. Finally, to model a very ordinal skewed distribution, I created a new 4—point scale, where 0 = 1; 1 and 2 = 2; 3, 4, and 5 = 3; and 6, 7, 8, 9, and 10 = 4. Again I computed Pearson and Spearman correlations and looked at the relation between the two (Table 1).

For the original data, the correlation between Spearman and Pearson coefficients was 0.99, the slope was 1.001, and the intercept was −0.007. Even with the severely skewed data, the correlation was still 0.987, the slope was 0.995, and the intercept was −0.0003. The means of the Pearson and Spearman correlations were within 0.004 for all conditions.

For this set of observations, the Pearson correlation and the Spearman correlation based on ranks yielded virtually identical values, even in conditions of manifestly non-normal, skewed data. Now it turns out that, when you have many tied ranks, the Spearman gives slightly different answers than the Pearson, but this reflects error in the Spearman way of dealing with ties, not a problem with the Pearson correlation. The Pearson correlation like all parametric tests we have examined, is extremely robust with respect to violations of assumptions.

4) You cannot use an intraclass correlation (or Generalizability Theory) to compute the reliability because the data are nominal/ordinal and you have to use Kappa (or Weighted Kappa)

Although this appears to be a special case of the previous section, there is a concise answer to this particular question. Kappa was originally developed as a “Coefficient of
agreement for nominal scales” (Cohen 1960), and in its original form was based on agreement expressed in a $2 \times 2$ frequency table. Cohen (1968) later generalized the formulation to “weighted kappa”, to be used with ordinal data such as Likert scales, where the data would be displayed as agreement in a $7 \times 7$ matrix. Weighting accounted for partial agreement (Observer 1 rates it 6; Observer 2 rates it 5). Although any weighting scheme is possible, the most common is “quadratic” weights, where disagreement of 1 unit is weighted 1, of 2 is weighted 4, of 3, 9, and so forth.

Surprisingly, if one proceeds to calculate an intraclass correlation with the same 7-point scale data, the results are mathematically identical, as proven by Fleiss and Cohen (1973). And if one computes an intraclass correlation from a $2 \times 2$ table, using “1” when there is agreement and “0” when there is not, the unweighted kappa is identical to an ICC. Since ICCs and G theory are much more versatile (Berk 1979), handling multiple observers and multiple factors with ease this equivalence is very useful.

Summary

Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of “coming to the wrong conclusion”. These findings are consistent with empirical literature dating back nearly 80 years. The controversy can cease (but likely won’t).

References


