When Does a Plan Differ Significantly from Other Plans? Statistical and Practical Significance Criteria of Star Ratings for the New Jersey Medicaid CAHPS Survey

Donna O. Farley, Ron D. Hays, Marc Elliott

DRU-1885-AHCPR

December 1997

Prepared for Agency for Health Care Policy and Research
The purpose of this memo is to examine the considerations involved in establishing health plan star ratings for the CAHPS Medicaid survey administered by the New Jersey Medicaid Office of Managed Care. Although CAHPS provides methods, formulas, examples, and recommendations, the choice of thresholds for establishing the star ratings ultimately rests in the hands of the Medicaid program. As stated by the New Jersey staff, the goal is to provide new Medicaid recipients with meaningful information on differences among health plans based on the survey results.

The issues addressed in this paper relate to setting thresholds for important differences among health plans in the CAHPS multi-item composites. Testing of the CAHPS overall ratings (e.g., overall rating of health plan) revealed that the number of stars a plan received was not sensitive to different thresholds. However, the same was not true for the multi-item composites. We begin with an overview of the CAHPS rating method and then discuss the New Jersey survey data.

The CAHPS Rating Approach

The CAHPS method for establishing the number of stars a health plan receives for each CAHPS quality dimension compares the average rating of the health plan to the overall average rating for all other health plans in the sample. The overall mean serves as the standard of comparison, so that new enrollees can see how each plan differs from its “peer group” -- either better or worse. Plans with ratings similar to the overall average of other plans are given 2 stars, those that rate significantly worse than average are given 1 star, and those that rate significantly better are given 3 stars. For determining the significance of differences in ratings, two aspects of measurement need to be considered:

- Statistical significance -- the extent to which the observed differences are real or just occurred by chance, and

- Practical (or substantive) significance -- the size of the difference in a plan’s rating from the overall average (the distance from the average)

For each type of significance, a threshold is set that represents the level beyond which differences in plan ratings are important enough to report in the consumer rating report. Each type of threshold is set by using options in the CAHPS program for calculating composite scores. For the statistical significance threshold, the CAHPS Handbook recommends using a threshold of p<0.05. This threshold indicates that a significant difference would have less than a 5 percent chance of occurring if the true rating was not different from the overall mean for all the other health plans. It is the most commonly used criterion for statistical significance.

CAHPS establishes a formula for determining practical significance in which the threshold is set as a specified fraction (“t”) of the distance from the overall mean to the nearest endpoint of a scale. Values of t can range between 0 and 1. For example, on a 10-point scale with a mean of 6, a threshold of 0.15 would identify a distance of 0.6 (or 0.15*[10-6] = 0.6). The value for t must be "calibrated" by the user to correspond with his or her sense of practical significance. CAHPS does not make any recommendation for the practical significance threshold, although the Handbook references a threshold of 0.25*(minimum distance) as an example.
Choosing a Practical Significance Threshold

To calibrate the threshold for practical significance the user should look at both (1) how much variation there is in the actual values of plan ratings on each composite and (2) what size differences in ratings are differentiated (given 1 or 3 stars) at different values of t (.05, .1, .2, etc.). Several factors should be considered to inform and guide the choice of threshold, which ultimately is a judgment call. These factors, summarized here, tend to argue for the acceptability of smaller thresholds for practical significance.

- **Sample size.** With large sample sizes, sometimes there is large statistical power to detect practically insignificant differences. This may not be an issue for New Jersey, however, given that most of the CAHPS applications cling tightly to the minimum sample sizes recommended in the Implementation Handbook.

- **Response patterns.** There is evidence in the literature of a general phenomenon of skewed responses on questions regarding satisfaction with health care, corresponding to a reticence to indicate dissatisfaction. This clustering of responses at the higher ends of the response options restricts the ability to obtain statistically significant differences.

- **Scale compression.** Mean scores on the 4-point scale derived from Always/Usually/Sometimes/Never responses "psychologically compress" the true degree of differences indicated because there is only a 3-point (4-1) spread between the endpoints. For calibrating the threshold, it is important to look at both the differences in the means for the 4-point scales and differences in the percentage of responses that fall into the top ("always," or 4-point) box for each plan.

  For example, if two health plans were rated 3.85 (for plan A) and 3.60 (for plan B) on an item that we agreed should happen all the time, this seemingly small difference would correspond to an "Always" response for 85% of patients in plan A but only for 60% of patients in plan B (assuming that the remaining patients responded "Usually" to questions in the composite). This means that more than twice as many patients in plan B as plan A report some lapses in this essential service.

- **Equivalence of thresholds.** One would naturally like to apply a threshold for a composite rating that is equivalent in strictness to thresholds for the single items comprising the composite. Applying the same value of the threshold parameter 't' as was used on the single items, however, does not achieve this objective. The reason is that the act of averaging items results in less variable behavior for the composite than exists in the individual items that constitute it. This makes the means of composites appear even more similar than they truly are than is the case for single items.

Ratings in the New Jersey CAHPS Medicaid Survey

This discussion of the New Jersey CAHPS survey ratings for the composite items looks at the star ratings obtained for different thresholds, the means and standard deviation of responses, and the percentage of responses in the "Always" category. Table 1 presents this information for all survey responses combined. The table highlights health plans and composites that vary most in the number of stars they receive, depending on the choice of values for t (0, 0.15, and 0.25). It lists the numbers of stars that plans would receive with a t = 0 threshold. *All* the plans with 1 and 3 stars would have 2 stars under a t = 0.25 threshold. Those that would have 2 stars under a
t = 0.15 threshold are noted with an "*" in the table and the information for them is highlighted with shading. Some observations regarding information provided are as follows:

- The use of a t = 0.25 threshold washes out all differences across plans, some of which clearly are practically significant based on differences in the percentage of “Always” responses to a composite.
- A t = 0.15 threshold tends to equate to a change of 6 to 7 percentage points in percentage of “Always” responses, so that differences of that magnitude or larger from the overall mean are given 1 or 3 stars.
- Using a t = 0.15 threshold instead of t = 0 affects only 3 composites—getting needed care, takes time and history, and customer service. For the “takes time and history composite,” only one health plan is affected by using a t = 0.15 threshold.

Table 2 is a tabulation of health plan stars in New Jersey for all users (both child and adult). It gives the number of plans that differ from the mean of the other plans for the three combinations of meaningful difference threshold.

**Calibrating Thresholds for Composites**

Although the issue of calibrating thresholds for composites is secondary to the information provided in Table 1 for selecting a threshold, it is useful to see how much of an effect the averaging of items has on threshold values. To correct for reductions in variability induced by averaging, one must multiply the threshold parameter “t” by a factor “s,” which ranges from 0 to 1. In other words, one must lower the threshold. For an intuitive understanding, s decreases as the number of items going into the composite increases, and it increases as the internal consistency of the items in the composite increases (as measured by inter-item correlation), all other things being equal.

In terms of coefficient alpha, 

\[ s = \sqrt{\frac{1}{k + (1-k)\alpha}} \]

where k is the number of items in the composite.

Now a composite of k items and coefficient alpha=alpha with threshold a*s will be equivalent to a single item with threshold a. So if alpha=0.70 and k=5, s=0.68 and a 0.17 threshold for the composite is equivalent to a 0.25 threshold for an item. The adjustments calculated for each composite in the New Jersey survey results are given in Table 3.
Table 1
Tabulation of Health Plan Stars in New Jersey for All Users (Child & Adult)

<table>
<thead>
<tr>
<th>Composite</th>
<th>Measure</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care need</td>
<td># stars-t=0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 3.08</td>
<td>Diff. Mean</td>
<td>-.15</td>
<td>+.16</td>
<td></td>
<td></td>
<td>+.07</td>
<td>-.09</td>
<td>+.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD = 0.90</td>
<td>% Always</td>
<td>49</td>
<td>43</td>
<td>59</td>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long waits</td>
<td># stars-t=0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 2.99</td>
<td>Diff. Mean</td>
<td>-.17</td>
<td>-.19</td>
<td>+.17</td>
<td></td>
<td>+.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD = 0.87</td>
<td>% Always</td>
<td>38</td>
<td>35</td>
<td>50</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicate</td>
<td># stars-t=0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 3.55</td>
<td>Diff. Mean</td>
<td>72</td>
<td>68</td>
<td>71</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+.10</td>
</tr>
<tr>
<td>SD = 0.64</td>
<td>% Always</td>
<td>69</td>
<td>65</td>
<td>73</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time/history</td>
<td># stars-t=0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 3.31</td>
<td>Diff. Mean</td>
<td>59</td>
<td>53</td>
<td>61</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+.14</td>
</tr>
<tr>
<td>SD = 0.81</td>
<td>% Always</td>
<td>60</td>
<td>58</td>
<td>60</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Office staff</td>
<td># stars-t=0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 3.54</td>
<td>Diff. Mean</td>
<td>71</td>
<td>67</td>
<td>76</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+.08</td>
</tr>
<tr>
<td>SD = 0.70</td>
<td>% Always</td>
<td>68</td>
<td>65</td>
<td>71</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer service</td>
<td># stars-t=0</td>
<td>2</td>
<td>1*</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean = 2.87</td>
<td>Diff. Mean</td>
<td>-.16</td>
<td>-.15</td>
<td>+.12</td>
<td>+.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+.23</td>
</tr>
<tr>
<td>SD = 0.85</td>
<td>% Always</td>
<td>41</td>
<td>34</td>
<td>40</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Number of stars for each plan is provided in cells above. The asterisk indicates that the health plan would have 2 stars if a 0.15 meaningful difference threshold is used. All plans would have 2 stars if a 0.25 threshold is used. Differences from the mean of other plans is indicated along with the overall mean for plans that change from 1 or 3 stars to 2 stars when a meaningful difference threshold is imposed. Entries in row 3 of each composite represent percentage of "always" (top box) response for each plan.
Table 2

Plans Differing from the Mean by Threshold Values

<table>
<thead>
<tr>
<th>Difference Threshold (t)</th>
<th>CAHPS Composite</th>
<th>0.00</th>
<th>0.15</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care need</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Long waits</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Communicate</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Time/history</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Office staff</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Customer service</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 3

Estimation of Proportion Reduction ("s") in Threshold Parameter and Composite Threshold Parameter Equivalent to 0.25 Single-Item Threshold

<table>
<thead>
<tr>
<th>Composite</th>
<th>Alpha</th>
<th>Estimate of Proportion Reduction</th>
<th>Composite Threshold Parameter*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Child</td>
<td>Adult</td>
<td>pooled</td>
</tr>
<tr>
<td>Care need (4 items)</td>
<td>0.48</td>
<td>0.61</td>
<td>0.55</td>
</tr>
<tr>
<td>Long wait (4 items)</td>
<td>0.66</td>
<td>0.63</td>
<td>0.65</td>
</tr>
<tr>
<td>Communication (4 items)</td>
<td>0.79</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>Time/history (2 items)</td>
<td>0.67</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>Office staff (2 items)</td>
<td>0.76</td>
<td>0.70</td>
<td>0.73</td>
</tr>
<tr>
<td>Customer service (3 items)</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
</tbody>
</table>

* Value of threshold parameter for composite that is equivalent to a 0.25 threshold for a single item.