EXECUTIVE SUMMARY

This report summarizes the approach that was used to establish performance categories for the ISAT science and social science tests at each of two grade levels at which ISAT is administered.

The development of useful and defensible standards and the execution of the standard-setting process itself are complex tasks. The State Board of Education relied on the contributions of many talented educators throughout Illinois to successfully accomplish the standard setting process.

Prior to the meetings of the standard-setting panels themselves, ISBE convened committees of curriculum experts to develop descriptions of student knowledge and skill levels that define the four performance categories: Academic Warning, Below Standards, Meets Standards, and Exceeds Standards. Educators throughout Illinois extensively reviewed these descriptions or definitions before sending them to the standard-setting panelists.

Panels of recognized subject matter experts subsequently convened in Springfield to translate the verbal definitions into cut scores on the ISAT tests (i.e., scores that define the boundaries between categories). Panelists were drawn from a pool of educators who had specific knowledge of student performance at the grade levels being assessed by ISAT and experience in assessing students at those grade levels. Panelists were selected to be broadly representative of the geographic and ethnic diversity of Illinois’ public school system. A total of 56 educators participated in the standard-setting process. The distribution across learning areas was as follows: science—28; social science—28.

A procedure originally proposed by Angoff is one of the most frequently used methods for determining cut scores when multiple-choice test scores are used. It can be most simply described as a focused, judgmental process by knowledgeable content experts.

In the most frequent application of the Angoff method (e.g., to establish a pass-fail standard), panelists are asked to examine an item and decide what proportion of minimally competent individuals will answer the question correctly. With respect to the ISAT, however, instead of being asked about minimally competent students, panelists were asked to indicate what percentage of three groups of students—those who were just above the Academic Warning/Below Standards boundary, those who were just above the Below Standards/Meets Standards boundary, and those who were just above the Meets Standards/Exceeds Standards boundary—would answer the question correctly. The ratings were made sequentially rather than simultaneously (i.e., panelists judged the proportion of correct responses by a criterion group to
every item before moving on to the next criterion group). Item performance
statistics were provided to help panelists anchor their ratings.

When the cut scores that emerged from this panel’s work were applied to score
distributions based on the 2000 ISAT census administrations, the percentages
of Illinois public school students in each science category were as shown in
the chart below.

As this chart shows, the percentages of students not meeting standards in
science ranges from 28% to 36%. The percentage of students who exceed
standards is larger at 7\textsuperscript{th} grade (18\%) compared to 13\% at 4\textsuperscript{th} grade.

The cut scores that emerged from the social science panel’s work lead to the
distributions shown in the next chart. The percentage of students not meeting
standards is virtually the same (41\%-42\%) at both grades. More students fall in
the meets category at 4\textsuperscript{th} grade (53\%) than at 7\textsuperscript{th} grade (46\%), but the
percentage of students in the exceeds category is greater at 7\textsuperscript{th} grade (12\%)
than at 4\textsuperscript{th} grade (6\%).

A number of checks were made on the adequacy of the ratings. Agreement
among the panelists was excellent at all grade levels. Evaluation forms com-
pleted at the end of the session indicated that the overall level of panelist confi-
dence in the ratings was extremely high.

These cutoffs represent a set of fixed benchmarks against which schools can
measure the success of their improvement efforts. The percentages of students
who fall into each category may shift each year in response to changes in the
student populations tested. However, as school improvement efforts are
effectively implemented throughout the state, the expectation is that these
percentages will systematically spiral upward.
1. STANDARD-SETTING PROCEDURES

Technical procedures by which standards are translated into test scores have been available for many years. They are most often applied in the case of tests used for certification or licensure when a cutoff must be defined that separates qualified from unqualified test takers. A procedure originally described by Angoff is most frequently used in such cases. This procedure can be most simply described as a focused, judgmental process by knowledgeable content experts. A modification of the basic Angoff procedure was used to establish cutoffs for the ISAT tests. In the most frequent application of this method (e.g., to establish a pass-fail standard), panelists are asked to decide what proportion of minimally competent individuals will answer each question correctly. With respect to the ISAT, however, instead of being asked about minimally competent students, panelists were asked to indicate what percentage of three groups of students—those who were just above the Academic Warning/Below Standards boundary, those who were just above the Below Standards/Meets Standards boundary, and those who were just above the Meets Standards/Exceeds Standards boundary—would answer the question correctly. The ratings were obtained sequentially. Panelists first made the Below Standards/Meets Standards cut for every item, then the Academic Warning/Below Standards cut, and, finally, the Meets Standards/Exceeds Standards cut. Item statistics were provided to help panelists anchor their ratings.

Participants

Panels of recognized subject matter experts convened for a two-day period during the week of April 10, 2000 in Springfield. Panelists were drawn from a pool of educators who had specific knowledge of students’ capabilities at the grade levels being assessed by ISAT and experience in assessing students at those grade levels. A single panel was used at each grade level.

The educators who served as panelists were drawn from schools and districts throughout the state. They were recruited by staff from ISBE’s Standards Division. The vast majority of panelists were classroom teachers. Participants were assigned to each panel on the basis of their background and experience. That is, fourth-grade teachers worked with the fourth-grade test, seventh-grade teachers with the seventh-grade test, etc. Panelists were selected to be broadly representative of the geographic and ethnic diversity of Illinois’ public school system. The distribution of panelists across grades and learning areas is shown in Table 1. A total of 56 educators established standards for the ISAT science and social science tests.

---

1 For simplicity, these cutoffs will be referred to subsequently as the “Meets” (Below Standards/Meets Standards), “Below” (Academic Warning/Below Standards), and “Exceeds” (Meets Standards/Exceeds Standards) cutoffs.
Table 1
Distribution of Panelists Across Grades and Learning Areas

<table>
<thead>
<tr>
<th>Grade</th>
<th>Science</th>
<th>Social Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>13</td>
</tr>
</tbody>
</table>

Prior to the meeting, panelists were sent background material to prepare them for the standard setting process. The material consisted of a description of the standard-setting process and the definitions that had been developed.

**Group Leaders**

The overall operation of the standard-setting panels was under the direction of MetriTech professional staff. In addition, four individuals who had extensive experience with the ISAT assessments and the content they covered were selected to serve as table facilitators or group leaders. They participated in a three-hour pre-session designed to familiarize them with the procedures to be used. In addition, staff from ISBE’s Standards Division were available throughout the work days.

**The Rating Process**

The work of each panel required two full days to complete. The program began with a two-hour, large-group presentation that oriented participants to the task and explained the procedures that would be followed. The rating procedure was described in detail and time was provided to answer all questions.

The panelists then broke up into grade-level groups. Each group was moderated by facilitators (group leaders) who were not themselves involved in making ratings. They began by leading panelists in an extended discussion of the category to be rated. The focus was on the performance definitions and the knowledge and skills that defined the borderline student.

When panelists had a clear picture of the type of student to be rated in mind, group leaders proceeded to the ratings themselves. They began with sample items that were used for training purposes only. In this way, test items to be rated were not viewed by the panelists prior to the beginning of the rating process itself.

Item analysis data was made available to help panelists understand the difficulty level of individual items and anchor their judgments. Random samples of approximately 16,000 students at each grade who had taken the tests in February 2000 were used for this purpose. Students were split into one
of five ability levels (quintiles) based on their performance on the overall test. That is, students whose overall scores placed them in the lowest fifth in terms of the score distribution (1-20%) represented the first group of students, students whose scores placed them in the next lowest fifth (21-40%) represented the second group of students, and so forth. For each group, panelists were shown the percentages of students within that group who had gotten the item correct.

For each item, panelists examined the item and the associated data, then made an initial rating. When all panelists had completed their initial rating, they reported their ratings to the group. If there was more than a 20% difference in the range of ratings, the group leader led a discussion of the ratings. The purpose of this discussion was not to force consensus but rather to allow the panelists to discuss the reasons for their ratings. This often resulted in one or more persons becoming aware of some facet of the item that they had not originally considered in their ratings. When the discussion was completed, panelists were asked to make a final rating on the item. Final ratings were not announced to the rest of the group.

After panelists completed their ratings of the sample items, group leaders passed out the actual 2000 ISAT test questions and rating forms. Initially, groups worked with one item at a time. As group members felt more comfortable with the process, they began rating several items prior to discussing them. All groups first completed ratings for the Meets cutoff, followed by ratings for the Below and then the Exceeds cutoffs.

The process proceeded cautiously and carefully. At the end of the first day of work, each group had completed ratings related only to the Meets cutoff.

The same procedure was followed on the second day to make the Below and Exceeds cutoffs. Groups began with a discussion of the level to be rated, worked with practice items, and then made their ratings of the actual test questions. Both sets of ratings were completed by the end of the second day.

When the ratings were finished, all panelists were asked to complete an evaluation form that was used to obtain their reactions to the procedure. Panelists also offered suggestions for enhancing the existing definitions and provided recommendations regarding specific ISAT items.

In some rating applications, panelists are asked to review an item and make multiple ratings simultaneously. On the surface this approach appears to be more efficient. However, the difficulty panelists have in developing and maintaining a clear image in their minds of the kinds of students they are rating and in shifting between reference groups suggest that this approach is likely to produce less reliable ratings.
RESULTS

Reliability of the Ratings

A number of checks were made on the ratings. First, the ratings at each grade level were analyzed to obtain the variance component estimates necessary to calculate generalizability coefficients. These coefficients, technically intraclass correlations, represent the degree of agreement among panelists in their ratings. Two types of coefficients\(^3\) are reported in Table 2. The first coefficient (“interrater”) represents the average level of correlation between the ratings of any two panelists. The second coefficient (“intergroup”) represents the level of correlation to be expected between the average ratings of one group of panelists and a second, similarly sized group of panelists.

On average, the interrater coefficient is .95 across the two areas and the intergroup coefficient is .99. There is no systematic difference in the reliability of the ratings across learning areas.

In terms of statistical indexes that may be more familiar, consider the following illustration. If the ratings were item scores and the items were persons (i.e., the kind of score matrix that is usually used for analyzing the reliability of a test), then the first intraclass correlation (interrater) would be interpreted as the average inter-item correlation, and the second intraclass correlation (intergroup) would be interpreted as coefficient alpha ($\alpha$). Thus, the reliability of these ratings exceed those of the best individual achievement tests.

As a check on the impact of group discussion on ratings, means and standard deviations for the average preliminary item ratings and average final item ratings were calculated. These are presented in Table 3. As this table shows, mean scores are remarkably consistent from initial to final rating period. Standard deviations tend to decrease between preliminary and final ratings. One interpretation of these data is that the discussion resulted in a tempering of extreme ratings, both high and low, toward the group mean. Group discussions did not, however, systematically raise or lower the final ratings.

\(^3\) The formulas used to calculate the two coefficients are as follows:

Intraclass correlation among $r$ ratings (interrater) = \(\frac{MS_r - MS_{pxr}}{MS_r + (r-1) MS_{pxr}}\)

Intraclass correlation of an average of $r$ ratings (intergroup) = \(\frac{MS_r - MS_{pxr}}{MS_r}\)

where $MS_r$ is the mean square between ratings and $MS_{pxr}$ is the mean square for persons by ratings.
Table 2
Generalizability Coefficients for Ratings

<table>
<thead>
<tr>
<th></th>
<th>SCIENCE</th>
<th>SOCIAL SCIENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interrater</td>
<td>Intergroup</td>
</tr>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below</td>
<td>0.92</td>
<td>0.99</td>
</tr>
<tr>
<td>Meets</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Grade 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Meets</td>
<td>0.90</td>
<td>0.99</td>
</tr>
<tr>
<td>Exceeds</td>
<td>0.95</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 3
Comparison of Initial and Final Ratings

<table>
<thead>
<tr>
<th></th>
<th>Initial Rating</th>
<th>Final Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>SCIENCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>14.44</td>
<td>0.61</td>
</tr>
<tr>
<td>Grade 7</td>
<td>25.09</td>
<td>1.09</td>
</tr>
<tr>
<td>Meets Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>36.67</td>
<td>1.80</td>
</tr>
<tr>
<td>Grade 7</td>
<td>33.52</td>
<td>2.42</td>
</tr>
<tr>
<td>Exceeds Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>58.44</td>
<td>0.61</td>
</tr>
<tr>
<td>Grade 7</td>
<td>55.18</td>
<td>0.35</td>
</tr>
<tr>
<td>SOCIAL SCIENCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Below Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>27.23</td>
<td>0.85</td>
</tr>
<tr>
<td>Grade 7</td>
<td>21.32</td>
<td>0.80</td>
</tr>
<tr>
<td>Meets Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>42.95</td>
<td>1.35</td>
</tr>
<tr>
<td>Grade 7</td>
<td>42.65</td>
<td>1.53</td>
</tr>
<tr>
<td>Exceeds Cutoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>62.65</td>
<td>0.78</td>
</tr>
<tr>
<td>Grade 7</td>
<td>58.44</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Checks were also made for potential outliers (i.e., raters who were unusually high or low relative to the ratings of the group as a whole). A few such ratings were found and removed from the final calculation of the cutoffs. However, removal of outliers did not systematically or significantly alter the cutoff scores.
**Cut Scores**

When panelist’s ratings are averaged across test items, the result is a “cut score” on the test that distinguishes two groups of examinees. For example, if panelists believe that 40% of borderline group students will answer each of 10 questions correctly, then a raw score of \(0.4 \times 10 = 4\) is the minimum score a student must obtain to be judged above the standard on this hypothetical 10-item test.

These raw scores were next converted to proficiency levels using results of the Rasch calibrations of each test. Finally, the scaling constants used to transform each theta value to the ISAT scale were applied to these values. The result was a scale score that represented the minimum acceptable scale score for entry into the category. These scale score values are shown in Table 4.

<table>
<thead>
<tr>
<th>SCIENCE</th>
<th>Academic Warning</th>
<th>Below Standards</th>
<th>Meets Standards</th>
<th>Exceeds Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>120-129</td>
<td>130-153</td>
<td>154-178</td>
<td>179-200</td>
</tr>
<tr>
<td>07</td>
<td>120-141</td>
<td>142-150</td>
<td>151-174</td>
<td>175-200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOCIAL SCIENCE</th>
<th>Academic Warning</th>
<th>Below Standards</th>
<th>Meets Standards</th>
<th>Exceeds Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>120-141</td>
<td>142-156</td>
<td>157-183</td>
<td>184-200</td>
</tr>
<tr>
<td>07</td>
<td>120-132</td>
<td>133-156</td>
<td>157-178</td>
<td>179-200</td>
</tr>
</tbody>
</table>

Table 5 shows the percentages of students meeting each standard based on the cutoff scores given in Table 3 for the total 2000 ISAT test populations.

**Participant Evaluations**

Evaluation forms were completed by each panelist at the end of the panelist’s two-day session. Three open-ended questions were asked of each participant: What were the most positive aspects of the experience? What aspects of the work caused you the most difficulty? What recommendations would you make for improving future sessions of this type? In addition, participants were asked three questions designed to assess their confidence in each of the three cutoffs and provided a 10-point scale on which to mark their answers. The lower end of the scale (1) was anchored by the phrase “Not very confident” and the upper end of the scale (10) was anchored by the phrase “Very confident.”

| Table 5 |
### Percentages of Illinois Students At Each Grade Level Who Fall Into Each Category

<table>
<thead>
<tr>
<th></th>
<th>Academic Warning</th>
<th>Below Standards</th>
<th>Meets Standards</th>
<th>Exceeds Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>1</td>
<td>35</td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td>Grade 7</td>
<td>12</td>
<td>16</td>
<td>54</td>
<td>18</td>
</tr>
<tr>
<td><strong>Social Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>11</td>
<td>30</td>
<td>53</td>
<td>6</td>
</tr>
<tr>
<td>Grade 7</td>
<td>3</td>
<td>39</td>
<td>46</td>
<td>12</td>
</tr>
</tbody>
</table>

Note: These percentages are based on data obtained from NCS in April 2000. There may be very slight differences in the percentages shown in the final school/district reports.

In terms of participants’ confidence in the Meets ratings, the average rating across all grades and all areas was 8.42, with a modal rating of 8.00. In terms of participants’ confidence in the Below ratings, the average rating across all grades and all areas was 7.77, with a modal rating of 9.00. In terms of participants’ confidence in the Exceeds ratings, the average rating across all grades and all areas was 8.35, again with a modal rating of 9.00. Participants in all grade levels and all areas expressed very high levels of confidence in their judgments of cutoff scores.