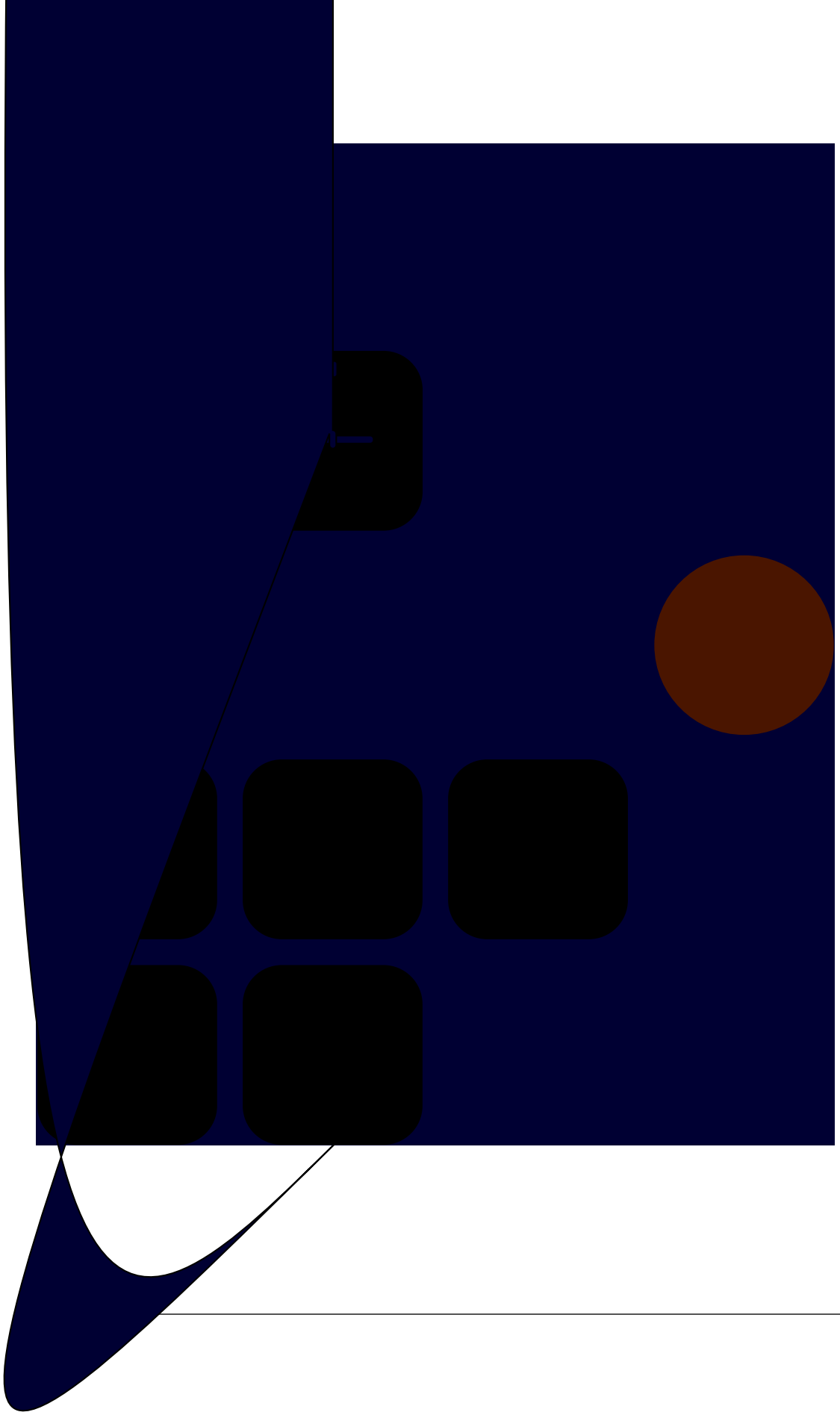


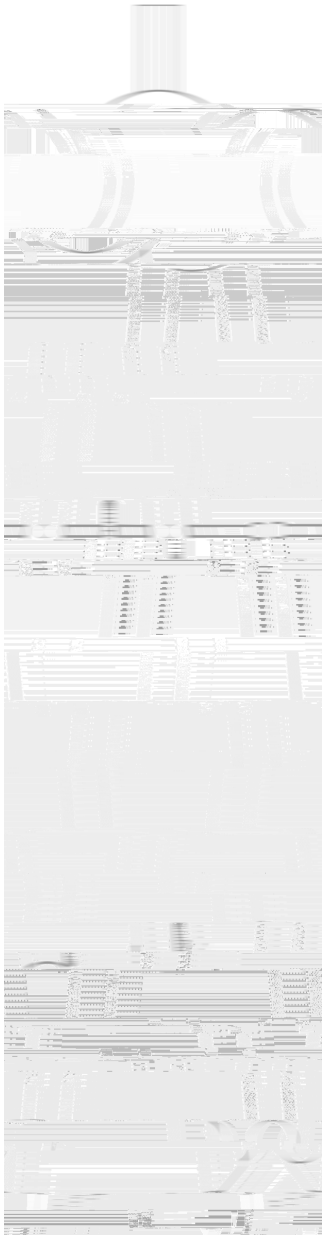


**Exhibit 12.1 Item Statistics for a Multiple-Choice Item - TIMSS 1999 Countries**

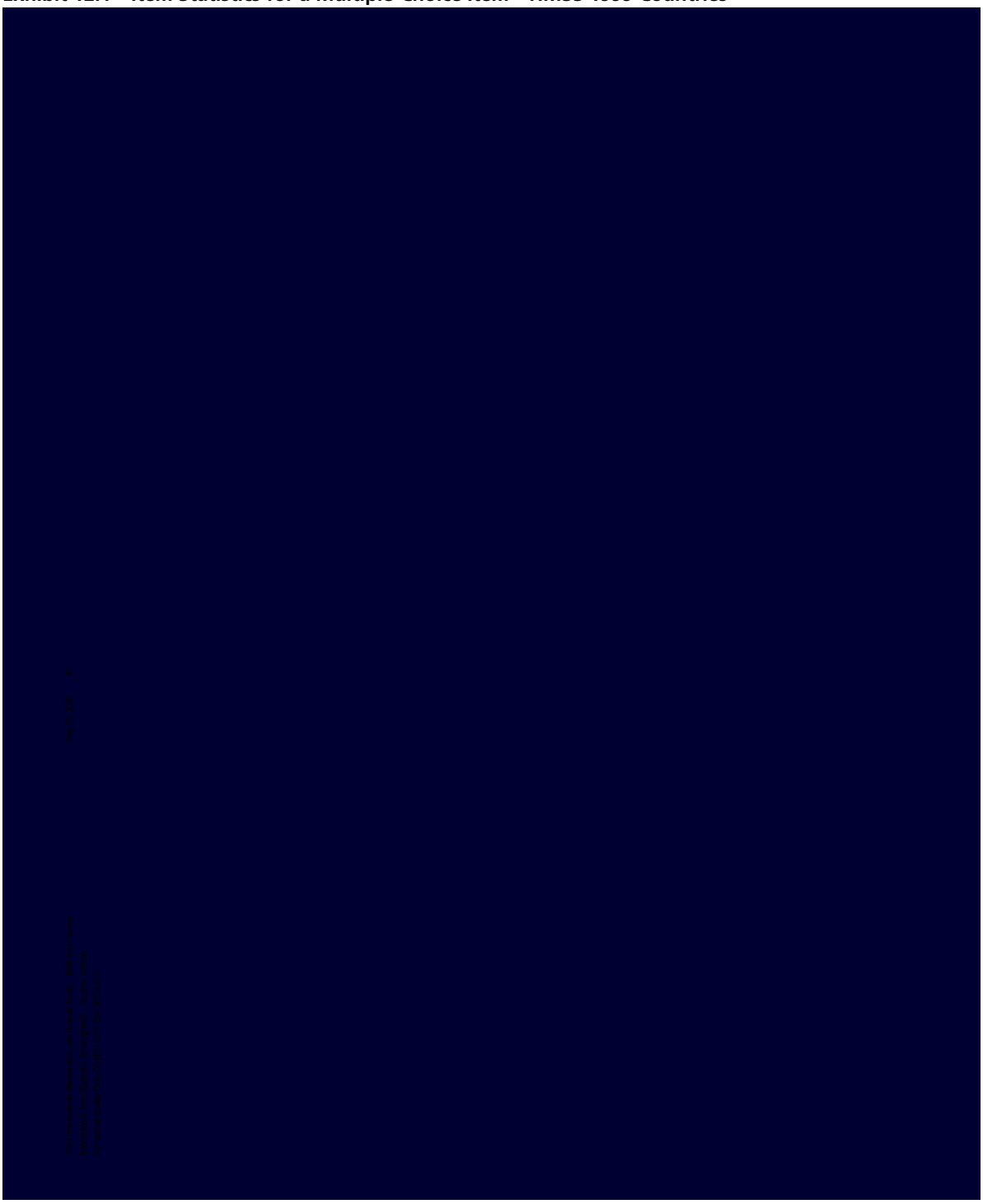


**Exhibit 12.2 Item Statistics for a Multiple-Choice Item - TIMSS 1999 Benchmarking Jurisdictions**

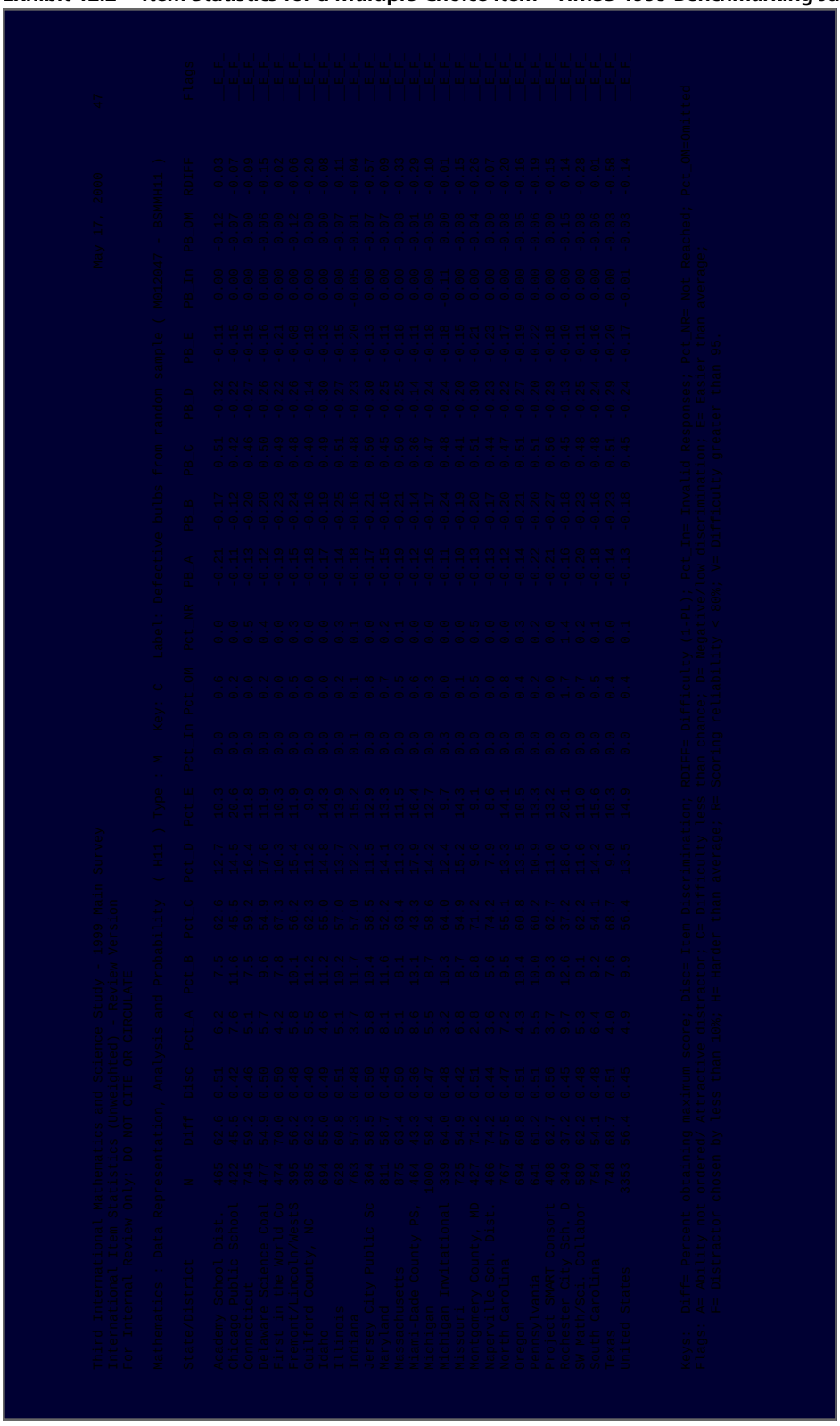


**Exhibit 12.3 Item Statistics for a Free-Response Item - TIMSS 1999 Countries**

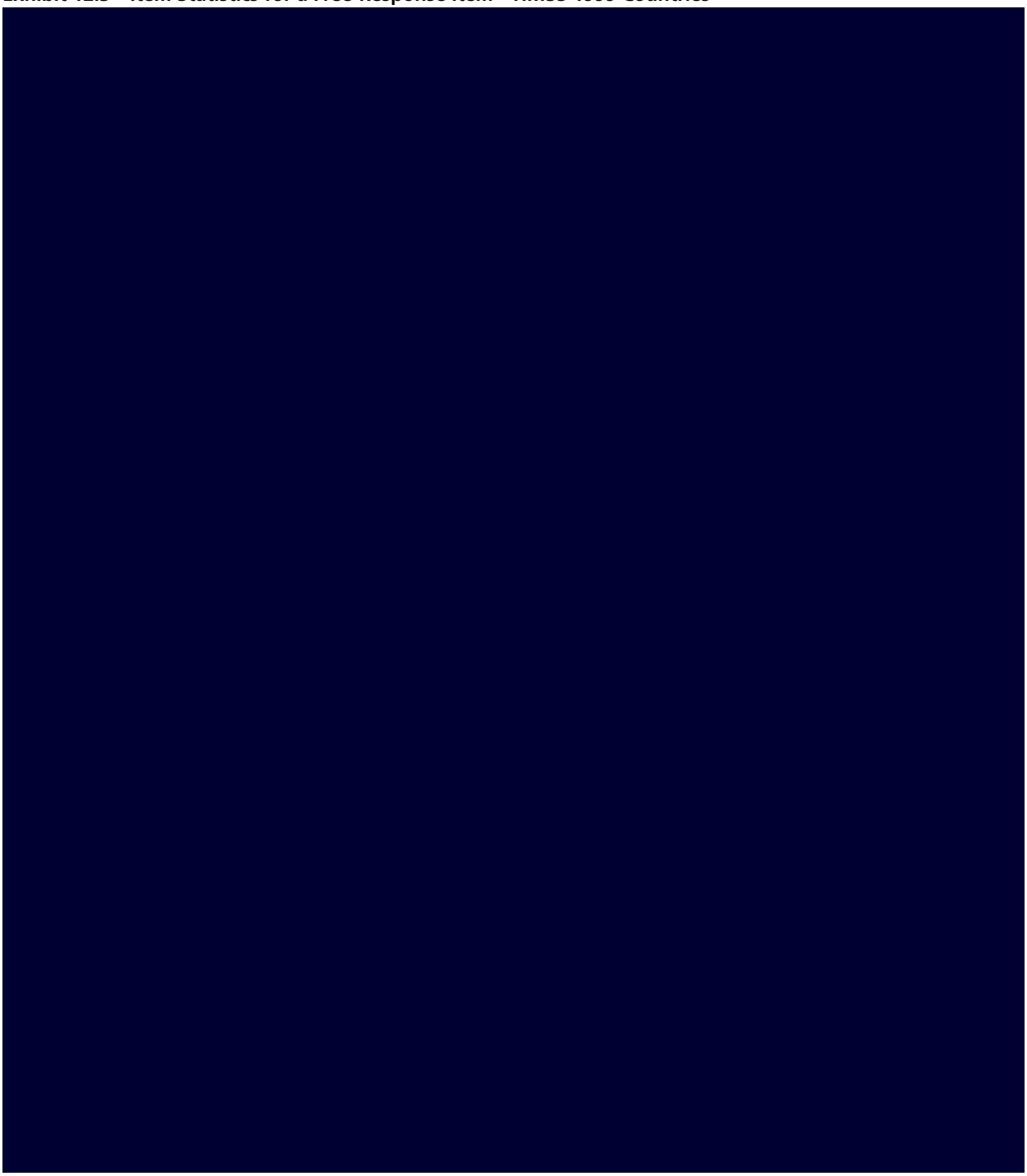
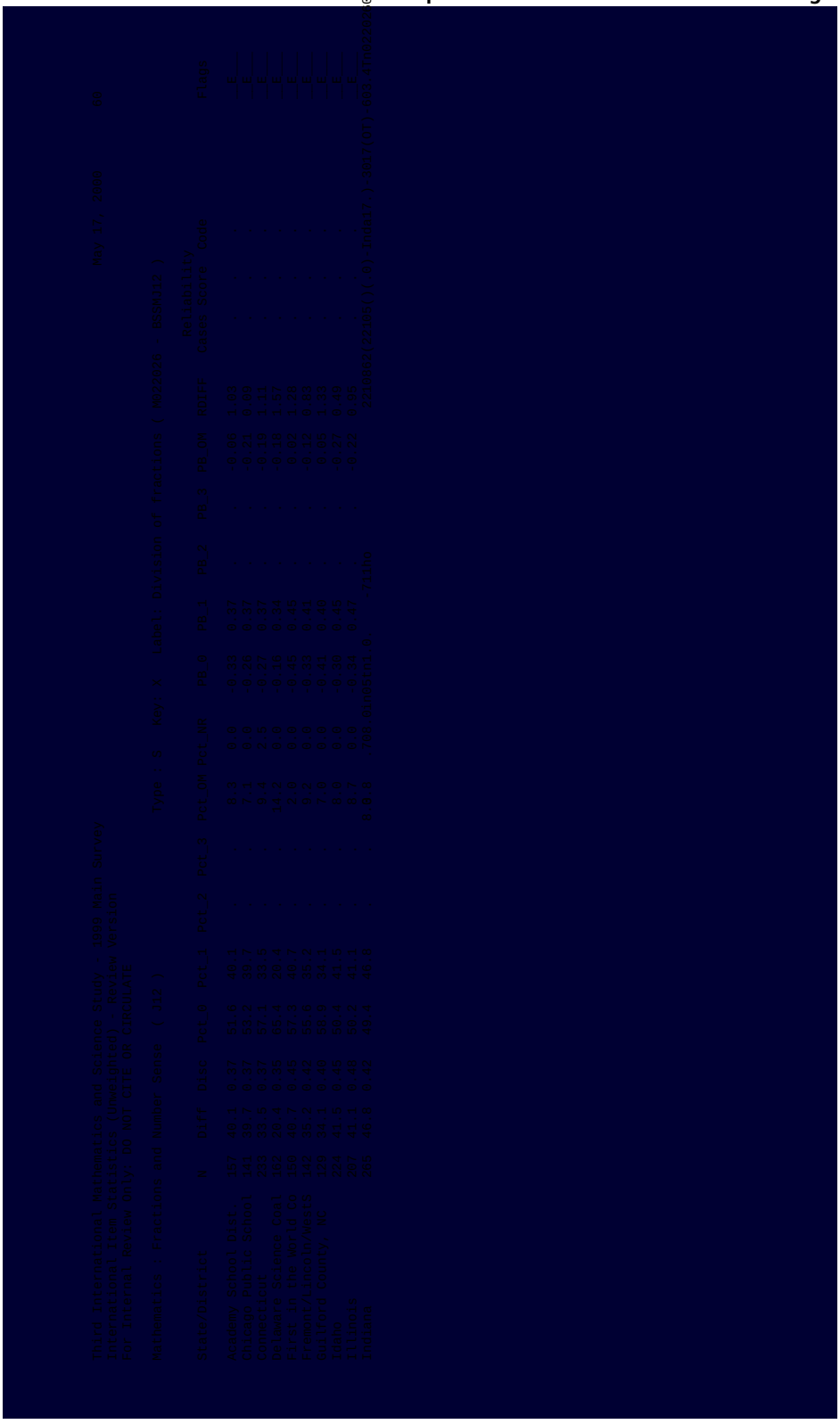


**Exhibit 12.4 Item Statistics for a Free-Response Item - TIMSS 1999 Benchmarking Jurisdictions**



Exhibits 12.1 through 12.4 contain the statistics described below.

**N:** This is the number of students to whom the item was administered. If an item was not reached by a student it was considered to be not administered for the purpose of the item analysis.[4]

**Diff:** The item difficulty is the percentage of students that provided a fully correct response to the item. In the case of free-response items worth more than one point this was the percentage of students achieving the maximum score on the item. When computing this statistic, items that were not reached were treated as not administered.

**Disc:** The item discrimination is the correlation between a correct answer to the item and the total score on all of the items in the subject area in the test booklet.[5] This correlation should be moderately positive for items with good measurement properties.

**PCT_A, PCT_B, PCT_C, PCT_D and PCT_E:** Used for multiple-choice items only (Exhibits 12.1 and 12.2), these represent the percentage of students choosing each response option for the item. Not reached items were excluded from the denominator for these calculations.

**PCT_0, PCT_1, PCT_2 and PCT_3:** Used for open-ended items only (Exhibits 12.3 and 12.4), these are the percentages of students scoring at each score level for the item. Not reached items were excluded from the denominator for these calculations.

**PCT_IN:** Used for multiple-choice items only, this is the percentage of students that provided an invalid response to a multiple-choice item. Invalid responses were generally the result of choosing more than one response option.

**PCT_OM:** This is the percentage of students that did not provide a response to the item even though the item was administered and they had reached it. Not reached items were excluded from the denominator when calculating this statistic.

○○○

4. In TIMSS, for the purposes of item analysis and item parameter estimation in scaling, items not reached by a student were treated as if they had not been administered. For purposes of estimating student proficiency, however, not reached items were treated as incorrectly answered.
5. For free-response items, the discrimination is the correlation between the number of score points and total score.

**PCT_NR:** This is the percentage of student that did not reach the item. An item was coded as not reached when there was no evidence of a response to any of the items following it in the booklet and the response to the item preceding it was omitted.

**PB_A, PB_B, PB_C, PB_D and PB_E:** Used for multiple-choice items only, these present the correlation between choosing each of the response options A, B, C, D, or E and the score on the test booklet. Items with good psychometric properties have zero or negative correlations for the distracter options (i.e., the incorrect options) and moderately positive correlations for the correct answer.

**PB_0, PB_1, PB_2 and PB_3:** Used for free-response items only, these present the correlation between the score levels on the item (zero, one, two, or three) and the score on the test booklet. For items with good measurement properties the correlation coefficients should change from negative to positive as the score on the item increases.

**PB_OM:** This is the correlation between a binary variable—indicating an omitted response to the item—and the score on the test booklet. This correlation should be negative or near zero.

**PB_IN:** Used for multiple-choice items only, this presents the correlation between an invalid response to the item (usually caused by selecting more than one response option) and the score on the test booklet. This correlation also should be negative or near zero.

**RDIFF:** This is an estimate of the difficulty item based on a Rasch one-parameter IRT model. The difficulty of the item is expressed in the logit metric (with a positive logit indicating a difficult item) and was scaled so that the average Rasch item difficulty was zero within each country.

**Reliability—Cases:** It was expected that the free-response items in approximately one-quarter of the test booklets would be scored by two independent scorers. This column indicates the number of times each item was double-scored in a country.

**Reliability—Score:** This column contains the percentage of times the two independent scorers agreed on the score level for the item.

**Reliability—Code:** This column contains the percentage of times the two scorers agreed on the two-digit code (i.e., score and diagnostic code) for the item.

As an aid to reviewers, the item-analysis display includes a series of "flags" signaling the presence of one or more conditions that might indicate a problem with an item. The following conditions are flagged:

- Item difficulty exceeds 95% in the sample as a whole
- Item difficulty is less than 25% for 4-option item as a whole

To help examine item-by-group interactions, the International Study Center produced a graphical display for each item showing the average probability across all groups of a correct response for a student of average international proficiency, compared with the probability of a correct response by a student of average proficiency in each group (see Exhibit 12.5 for an example). The probability for each group is presented as a 95% confidence interval, which includes a built-in Bonferroni correction for multiple comparisons.

The limits for the confidence interval are computed as follows:

where $RDIFF_{ik}$ is the Rasch difficulty of item $k$ within group $i$; $SE_{RDIFFik}$ is the standard error of the difficulty of item $k$ in group $i$; and $Z_b$ is the critical value from the $Z$ distribution, corrected for multiple comparisons using the Bonferroni procedure.

## 12.3 Item Checking Proceduresmultiple (documragoÞcietTj /F8 1 translo Þcietveri Þho Þcie0 0 6.601 254.761 414.74

# References