Table S1

*Average SE of Human- and Machine-Coded New Science Item Parameter Estimates*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Human-Coded New Science Items |  | Machine-Coded New Science Items |
| *n* | FIPC-P | FIPC-S | CC-S |  | FIPC-P | FIPC-S | CC-S |
| *SE of Slope Parameter Estimates* |
| 125 | 0.225 | 0.207 | 0.200 |  | 0.218 | 0.190 | 0.184 |
| 250 | 0.158 | 0.145 | 0.141 |  | 0.151 | 0.134 | 0.130 |
| 500 | 0.110 | 0.100 | 0.100 |  | 0.105 | 0.095 | 0.089 |
| 1,000 | 0.077 | 0.071 | 0.071 |  | 0.077 | 0.063 | 0.063 |
| *SE of Difficulty Parameter Estimates* |
| 125 | 0.202 | 0.202 | 0.202 |  | 0.195 | 0.195 | 0.195 |
| 250 | 0.141 | 0.141 | 0.145 |  | 0.138 | 0.138 | 0.138 |
| 500 | 0.100 | 0.100 | 0.100 |  | 0.095 | 0.095 | 0.095 |
| 1,000 | 0.071 | 0.071 | 0.071 |  | 0.071 | 0.071 | 0.071 |
| *SE of Step Parameter Estimates* |
| 125 | 0.230 | 0.230 | 0.232 |  | 0.249 | 0.249 | 0.249 |
| 250 | 0.161 | 0.161 | 0.161 |  | 0.160 | 0.160 | 0.173 |
| 500 | 0.114 | 0.114 | 0.114 |  | 0.118 | 0.118 | 0.118 |
| 1,000 | 0.077 | 0.077 | 0.077 |  | 0.083 | 0.083 | 0.083 |

***Note***. $\overbar{n}$ = country-specific sample size; SE = standard error; standard errors in parentheses.

Table S2

*Average MD of Science items across calibration methods and subsamples*

|  |  |  |  |
| --- | --- | --- | --- |
|  | New Science Items |  | Trend Science Items |
| *n* | FIPC-P | FIPC-S | CC-S |  | FIPC-P | FIPC-S | CC-S |
| 125 | 0.000(0.0002) | 0.000(0.0002) | 0.000(0.0002) |  | 0.001(0.002) | 0.000(0.0004) | 0.000(0.0003) |
| 250 | 0.000(0.0002) | 0.000(0.0002) | 0.000(0.0002) |  | 0.001(0.002) | 0.000(0.0003) | 0.000(0.0002) |
| 500 | 0.000(0.0002) | 0.000(0.0002) | 0.000(0.0002) |  | 0.000(0.002) | 0.000(0.0003) | 0.000(0.0002) |
| 1,000 | 0.000(0.0002) | 0.000(0.0002) | 0.000(0.0002) |  | 0.000(0.002) | 0.000(0.0002) | 0.000(0.0002) |

***Note***. $\overbar{n}$ = country-specific sample size; MD = mean deviation; standard errors in parentheses.



*Figure S1*. Pairwise contrasts of parameter constraints (different amount of information introduced into the calibration) in terms of their estimated marginal mean bias for each sample size (overall new Science items). Note: Left column: item discriminations. Right column: item difficulties. The blue bars represent the 95% CIs of the means. Non-overlapping red arrows within the blue bars indicate substantial differences between parameter constraints. Contrasts conducted with Tukey multiplicity adjustment.$ α=0.05$.



*Figure S2*. Pairwise contrasts of parameter constraints (different amount of information introduced into the calibration) in terms of their estimated marginal mean bias for each sample size (human-coded new Science items). Note: Left column: item discriminations. Right column: item difficulties. The blue bars represent the 95% CIs of the means. Non-overlapping red arrows within the blue bars indicate substantial differences between parameter constraints. Contrasts conducted with Tukey multiplicity adjustment.$ α=0.05$.



*Figure S3*. Pairwise contrasts of parameter constraints (different amount of information introduced into the calibration) in terms of their estimated marginal mean bias for each sample size (machine-coded new Science items). Note: Left column: item discriminations. Right column: item difficulties. The blue bars represent the 95% Cis of the means. Non-overlapping red arrows within the blue bars indicate substantial differences between parameter constraints. Contrasts conducted with Tukey multiplicity adjustment.$ α=0.05$.