

# **The Impact of Differential Investment of Student Effort on the Outcomes of International Studies**

Jayne Butler

*University of Melbourne*

Raymond J. Adams

*Australian Council for Educational Research*

*University of Melbourne*

International comparative assessments of student achievement, such as Trends in Mathematics and Science (TIMSS) and Programme for International Student Achievement (PISA) are becoming increasingly important in the development of evidence-based education policy. The potentially far-reaching influence of such studies underscores the need for these assessments to be valid and reliable. In education, increasing recognition is being given to motivational factors which impact on student learning. This research considers a possible threat to the validity of such studies by investigating the influence the amount of effort invested by test-takers has on their outcomes. Reassuringly, it is found that the reported expenditure of effort by students is fairly stable across countries. This finding counters the claim that systematic cultural differences in the effort expended by students invalidate international comparisons. Realistically reported effort expenditure is related to reading achievement with an effect size similar to variables such as single parent family structure, gender and socio-economic background. Finally, when reporting trends, taking effort into account should be considered and may well facilitate the interpretation of national and gender trends in reading achievement.

## Introduction

In the popular literature there is widespread concern that assessments which have no direct consequences for students, teachers or schools underestimate student ability, and this underestimation increases as the students become even more familiar with the tests (Holliday and Holliday, 2003). This issue is particularly relevant for international comparative studies such as TIMSS and PISA. A number of commentators have argued that unknown differences across regions and cultures in test compliance and test motivation pose a serious threat to the validity of assessments particularly when results are to be compared across regions, time or cultures (Bracey, 1999; Holliday and Holliday, 2003). This potential problem of varying test compliance across countries participating in international studies is raised particularly when the achievement of students from a given country does not measure up to expectations.

In the academic literature there has been some research on the impact of student motivation on their test performance but this research has been limited and the outcomes somewhat equivocal. Debate has centred on whether students undertaking low-stakes tests like international comparative studies and pilot assessment programs may experience low levels of motivation (Wise and DeMars, 2005). Because of the personally non-consequential nature of the assessment, students may be demonstrating less than optimal levels of achievement. As a consequence it has been suggested that the overall results may provide an underestimation of student knowledge and proficiency and hence a threat to the validity of test scores and their interpretation (Kiplinger and Linn, 1996; Mislevy, 1995; O'Neil, Sugrue, and Baker, 1996; Wainer, 1993; Wolf and Smith, 1995).

The study of test-taking motivation has emerged as a research topic located within the attribution theory of motivation. Attribution theory is a social cognitive theory of motivation developed by Bernard Weiner in the 1980s. In this model cognitions (attributions) are hypothesized as pivotal determiners of affect. A key assumption of the Weiner model is that emotions are dependent on a cognitive appraisal process so the cen-

tral theme of the theory is the way an individual attempts to determine the cause of an event.

Weiner identified ability, effort, task difficulty, and luck as the most important factors affecting attributions for achievement. Attributions are classified along three causal dimensions: locus of control, stability, and controllability. The locus of control dimension has two poles: internal versus external locus of control. The stability dimension captures whether causes change over time or remain constant. Controllability ranges from causes an individual can control, such as skill or efficacy, to causes an individual cannot control, such as aptitude, mood, the actions of others, and luck (Weiner, 1986).

The amount of effort a person will expend in undertaking an activity is determined by an individual's perceptions or attributions for success or failure. Effort is internal and unstable but remains a factor over which the individual can exercise a great deal of control. Therefore, the basic principle of attribution theory as it applies to motivation is that an individual's own perception or attributions for success or failure determine the amount of effort the individual will expend on that activity (Weiner, 1986).

Motivation employed in taking a test is defined as a specific form of motivation. Test-taking motivation is a context dependent trait that can be defined as the individual's perceived motivation to do his or her best on a given test (Pintrich and Schunk, 2002). Motivational studies identify various indicators of motivation such as task choice, effort, perseverance and achievement. Despite concern about the role of student motivation, few studies have investigated the test-taking construct and its relation to achievement in the large-scale testing setting (Baumert and Demmrich, 2001).

In investigating this construct some researchers have chosen to examine the relationship between test-taking motivation and incentives. O'Neil, Sugrue, Abedi, Baker, and Golan (1997) showed in experimental studies on test motivation that a monetary incentive paid for each correct item enhanced the reported level of effort and test performance amongst grade 8 students. Among

grade 12 students, no effect of a monetary incentive on either reported level of effort or test performance was found. In a subsequent study on released TIMSS items, no experimental effect was found on performance amongst grade 12 students who were offered a monetary incentive (O'Neil, Abedi, Lee, Myoshi and Mastergeorge, 2004). However, these students reported a higher level of invested effort compared to the students in the control situation.

The release of the PISA 2000 results in Germany generated wide ranging discussions within the educational community. Articles written by Keitel and Kilpatrick, and Haenisch in 1998 (see Baumert and Demmrich, 2001, p. 442) made the claim that German students do not take achievement tests seriously unless their performance is graded. Baumert and Demmrich (2001) countered this claim by asserting that the conclusion drawn in the articles was not based on empirical evidence. In an experimental study of 467 students from Grade 9, Baumert and Demmrich found that the impact of test treatment conditions of informational feedback, grades, and financial incentives had no effect on intended and invested effort in test performance.

Overall, the findings related to test-taking motivation and achievement, reported from both field and laboratory studies, are contradictory. Research on the impact of incentives on student effort in test-taking settings show divergent results. Even within low-stakes contexts, the relationship between test-taking motivation and test achievement is not clear. Some studies have found that students are reasonably motivated to do their best even when the test is low stakes (Baumert and Demmrich, 2001). Other studies have found that the stakes of the test are indeed related to motivation and performance (Wolf and Smith, 1995).

The experimental findings related to test-taking motivation, patterns of gender and achievement are also contradictory. Studies investigating gender effects have not produced uniform findings. A study by Karmos and Karmos (1984) showed that the relationship between test-taking motivation and performance is stronger for boys

than girls. In contrast, Brown and Walberg (1993) found no interaction between test-taking motivation and gender.

Many aspects related to the perception of the stakes of the assessment by the student need to be considered. Our research explorations are founded on the notion that students are participants in the assessment process and bring knowledge and past experience to these settings. A goal of the present study is to explore the impact that student reported effort has on the outcomes of assessments. Another focus area is patterns of gender difference observed in the expenditure of effort within a low-stakes environment.

Data from two waves of PISA assessments, PISA 2000 and PISA 2003, are explored to establish the relationships between student effort and test outcomes. The research focus is on the extent to which key results, for example reading performance differences between countries and between subgroups within countries, might be influenced by the differential effort that students from varied backgrounds might invest in their performance on such tests.

### Measuring Effort

An overview of PISA, including its methods, products and purposes are provided in Turner and Adams (2007). Here we do, however, describe in some detail the instrument that is used to quantify test-taking motivation - the Effort Thermometer which was developed by a group of researchers based at the *Max-Planck-Institut* in Berlin (Kunter *et al.*, 2002)

In 2000, motivated by concern that students did not invest effort in large-scale studies such as PISA, three countries (Germany, Australia and Norway) included a rating of the amount of effort that students expended on the assessment. In 2003 the use of the scales was made an integral part of PISA and was administered to all participants.

#### *Effort Thermometer instructions*

The Effort Thermometer is administered at the end of the two-hour PISA test session. Students are given the brief instructions appearing in the text box below.

Please stop.  
 Now turn to the last page or so in your booklet, where there is a question about calculator use and a question about effort. Please answer these now, and then close your booklet.

The Effort Thermometer is based on three 10-point scales and is displayed in Figure 1. These scales are named High Personal Effort, PISA Effort and School Mark Effort. The first scale indicates the maximum effort invested in a situation that is of high personal importance to the participant. The second scale presents an opportunity to rate the effort expended in PISA. The third scale shows the anticipated expenditure of effort if the assessment were to have high personal relevance for the participant within the school context.

Figure 1 shows that optimum effort is indicated by a cross next to the number 10 on the scale. The three scales are indicated by a line of numbered boxes. Therefore a student may show maximum effort by crossing the box next to the number 10.

The scenarios described as part of the Effort rating scales and the Effort Thermometer instructions are not read to the students. It can be argued that

the Effort Thermometer could present a challenging reading task for a low ability reader. The scenario and instructions consist of four complex sentences; the image of the thermometer does not provide detailed decoding support for students unfamiliar with the word; finally the rating scales are not presented horizontally which is the more usual array for these types of rating scales but vertically to match the metaphor of the thermometer. Conceptually and linguistically the poor reader may be unable to show effort expenditure validly or may be deterred from responding to these scales.

*Initial instructions for students*

At this point, it is also worthwhile examining the instructions given to the students to see if they provide any motivational impetus for maximizing individual effort. The initial instructions do include some key information pertaining to the scope and purpose of the assessment. However, the instructions are brief and cannot be regarded as a motivational speech.

*How much effort did you invest?*

*Please try to imagine an actual situation (at school or in some other context) that is highly important to you personally, so that you would try your very best and put in as much effort as you could to do well.*

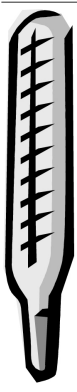
In this situation you would mark the highest value on the "effort thermometer" as shown below:	Compared to the situation you have just imagined, how much effort did you put into doing this test?	How much effort would you have invested if your marks from the test were going to be counted in your school marks?
 <input checked="" type="checkbox"/> 10	<input type="checkbox"/> 10	<input type="checkbox"/> 10
<input type="checkbox"/> 9	<input type="checkbox"/> 9	<input type="checkbox"/> 9
<input type="checkbox"/> 8	<input type="checkbox"/> 8	<input type="checkbox"/> 8
<input type="checkbox"/> 7	<input type="checkbox"/> 7	<input type="checkbox"/> 7
<input type="checkbox"/> 6	<input type="checkbox"/> 6	<input type="checkbox"/> 6
<input type="checkbox"/> 5	<input type="checkbox"/> 5	<input type="checkbox"/> 5
<input type="checkbox"/> 4	<input type="checkbox"/> 4	<input type="checkbox"/> 4
<input type="checkbox"/> 3	<input type="checkbox"/> 3	<input type="checkbox"/> 3
<input type="checkbox"/> 2	<input type="checkbox"/> 2	<input type="checkbox"/> 2
<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 1

Figure 1. The Effort Thermometer

The instructions in the box below were read to every participating student in 2003 before starting the two-hour PISA test session. Their goal is to provide basic background details relating to the assessment.

**Results**

The remainder of this paper is organized into four sections. Firstly, the construction of a comparable effort variable is described and explored across countries. The second section investigates the relationship between Relative Effort and reading achievement. The third section explores the relationship between Relative Effort and gender at the global and national levels. The fourth section concentrates on empirical analysis and interpretation of two waves of reading achievement data for students in Germany and Australia.

*Exploring the effort variable*

Preliminary analysis emphasized descriptive statistics for PISA Effort and School Mark Effort, followed by an investigation of the effort variables and reading achievement. The PISA 2003 data set was examined in these exploratory analyses.

*Constructing the effort variable*

For use in subsequent analyses we constructed a new variable called Effort Difference as follows:

**Effort Difference = PISA Effort – School Mark Effort**

The use of Effort Difference, rather than PISA Effort, as a key independent variable allows the investigation of how seriously students are viewing the PISA test compared to other aspects

of their school work which carry consequential outcomes for them. The construction of Effort Difference may also better compensate for cultural variations that may be activated when making effort ratings.

There are two reasons why the use of effort as a difference value is preferred over effort as an absolute value. Firstly, while the Effort Thermometer endeavours to ensure cultural comparability by setting up the High Personal Effort scale as an anchor the direct comparability of effort expenditure across cultures might still be questioned.

Secondly, it is possible that the yield of education systems where students are not as motivated or effortful may well be lower than the yield of education systems where students are more motivated and effortful. Therefore, finding that PISA Effort is lower may not be a finding that suggests that country performance differences are caused by effort differences, but rather they are co-varying outcomes of the education systems or cultures. Using the difference between PISA and School Mark Effort therefore is a way of neutralizing this component of differing effort within the school context.

The Effort Difference scores can range from negative nine to positive nine. Table 1 displays the (weighted) percentages of students from PISA 2003 scoring in each category of Effort Difference, along with the mean reading proficiency for students in each category.<sup>1</sup> The percentage of students who did not respond to the Effort Thermometer was 17.5 percent. It is

<sup>1</sup> The values reported throughout the paper were computed using appropriate weights and plausible values, as described in the PISA technical documentation (OECD, 2005).

You have been chosen to take part in an important international education study. This study is called the Programme for International Student Assessment, 'PISA' for short. Its goal is to find out what students your age all around the world know about reading, mathematics and science. There are about <number of> students representing <country>. Around the world there are about 200,000 students involved, from more than 7000 schools in 40 countries.

The results of the study will help countries determine what students are learning. Because the study may affect students all over the world in the future, we ask that you do the very best that you can.

possible that the magnitude of the omit rate could be influenced by Effort Thermometer readability issues highlighted earlier.

A negative score on Effort Difference means that students indicate they would try harder on a test that *counts* than they did on the PISA assessment. These students are making a realistic judgement about the amount of effort that is applicable for these types of low-stakes assessments.

Table 1 shows that most of the students (92.8%) who did respond to the Effort Thermometer did so in a predictable and realistic way. The percentage of students with positive Effort Difference scores, which would appear to be an unrealistic appraisal, constitutes 5.9% of the weighted sample, 7.2% of the valid responses. The score category zero, which includes students who indicated that they put an equal amount of effort into PISA compared to a test that counts towards their school marks, attracted 22.9 % of the weighted sample, 27.7% of the valid responses.

Table 1 also shows that for all positive values of Effort Difference the mean reading achievement is lower compared to students who responded in the more reasonable fashion.

The Effort Difference variable has two undesirable characteristics with respect to convenient data analysis and interpretation. First, the scores above zero have very low frequencies and second, the majority of the students have negative scores. For these reasons, two recodes have been implemented. The first recode entailed

collapsing all of the categories with positive values into a single category with score negative ten. This recode was based on the observation that students who provided an unrealistic response were low in number and exhibited low levels of reading ability. The second recode involved adding ten to the Effort Difference scores to facilitate description of the construct and interpretation of the findings from the analysis. The variable referred to as Effort Difference when recoded is subsequently referred to as Relative Effort.

Table 2 shows the percentage of students and mean reading proficiency scores for each category of Relative Effort, with the collapsed category labelled as score zero. Score ten becomes the modal category and includes those students expending equal amounts of effort on both the consequential and non-consequential tests.

To aid in the description and discussion of effort it is convenient to provide descriptors for five of the levels of Relative Effort. Students in category zero are labelled *unrealistic raters*; category one *PISA cynics*; category eight, *PISA realists*; category nine, *diligent realists*; and category ten, *PISA supporters*.

The values reported in Table 2 show that *unrealistic raters* (students who reported that they put more effort into PISA than they would put into a test that counted towards their schools marks) obtain the lowest mean for reading. Although not reported here, this low pattern of achievement is also observed for the domains of mathematical and scientific literacy.

Table 1

*Percentage of students in all categories of Effort Difference using weighted values for PISA 2003*

Effort Difference score	% of students	Mean achievement	Effort Difference score	% of students	Mean achievement
Score -9	0.5	437.34	Score 1	3.2	399.68
Score -8	0.3	456.14	Score 2	1.4	404.67
Score -7	0.6	463.38	Score 3	0.6	396.47
Score -6	1.0	463.14	Score 4	0.3	399.62
Score -5	2.7	470.19	Score 5	0.2	395.35
Score -4	3.8	480.60	Score 6	0.1	399.59
Score -3	8.5	488.71	Score 7	0.0	417.37
Score -2	16.5	490.84	Score 8	0.0	364.68
Score -1	19.7	489.83	Score 9	0.1	423.15
Score 0	22.9	464.23	Non-respondents	17.53	389.66

*PISA supporters* (scorers of ten on Relative Effort) are the largest group of students who express that they put an equal amount of effort into PISA compared to a test that *counts* (see Table 2).

*Diligent realists* (scorers of nine on Relative Effort) are the group of students with the second highest mean reading achievement in PISA.

*PISA realists* (scorers of eight on Relative Effort) are the group of students with the highest mean reading achievement in PISA.

*PISA cynics* (scorers of one on Relative Effort) indicate that they are putting very little effort into PISA. This group of students has the second lowest mean reading achievement.

In summary, the exploration of the effort variable has resulted in three actions. Firstly, the unrealistic ratings have been collapsed into a single group; secondly the variable Relative Effort has been rationalised and defined, and thirdly, labels for certain categories of raters on the Relative Effort scale have been adopted to aid description and explanation.

*Exploring effort and reading*

The first step in the global investigation of effort is to examine Relative Effort by country and subsequently relate this to the country’s performance in Reading for each type of effort variable. The means for PISA Effort, School Mark Effort and Relative

Table 2

*Percentage of students in recoded Relative Effort for PISA 2003*

Relative Effort score	% of students	Mean achievement	Relative Effort labels
score 0	5.9	401.54	Unrealistic raters
score 1	0.5	437.34	PISA cynics
score 2	0.3	456.14	
score 3	0.6	463.38	
score 4	1.0	463.14	
score 5	2.7	470.19	
score 6	3.8	480.60	
score 7	8.5	488.71	
score 8	16.5	490.84	PISA realists
score 9	19.7	489.83	diligent realists
score 10	22.9	464.23	PISA supporters
Non- respondents	17.5	389.66	
Total	100	459.58	

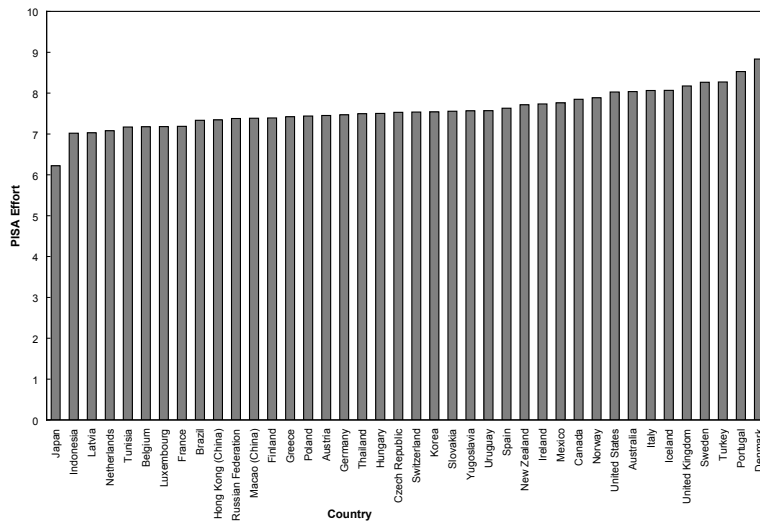


Figure 2. PISA Effort by country

Effort for each PISA 2003 country are shown in Figure 2, Figure 3 and Figure 4 respectively.

Figure 2 shows that typically countries score between seven and eight on PISA Effort. Japan has the lowest rating for PISA Effort while Thailand has the highest rating. The five countries scoring above eight are less economically developed countries. This result is consistent with previous research that shows that respondents from less economically developed countries tend to respond in a socially desirable fashion to scales of this type (King, Murray, Salomon, and Tandon, 2004). A difference of 2.61 is observed between the highest and lowest rating countries on PISA Effort. The difference across countries on reported levels of PISA Effort is relatively small, which tends not to support anecdotal evidence that students in some countries are not motivated to perform as well as those in other countries (Wainer, 1993).

Figure 3 shows that Japan has the lowest average rating for School Mark Effort at 8.13 points. Contrastingly, Denmark has the highest rating at 9.64 points. The overall difference of School Mark effort is 1.51 points. The larger difference between countries for PISA Effort could be a reflection of the status and lack of personal consequence of PISA.

Figure 4 shows Indonesia has the lowest mean for Relative Effort (6.49) and Thailand (8.47) has the highest. For most countries the means for Relative Effort are fairly consistent at about eight points. This corresponds to students on average saying that their PISA Effort is about two thermometer points below their School Mark Effort.

In these global rankings, Japan is in 14<sup>th</sup> place for mean reading achievement (OECD, 2004b) and in 40<sup>th</sup> place for mean Relative Effort. However, in terms of effort expenditure, Japanese students respond conservatively on both the PISA and School Mark Effort scales. Japanese students average 6.22 for PISA Effort and 8.13 for School Mark Effort. Overall Japanese students give low ratings on both the scales as compared to students in other countries. This pattern contrasts with Turkish students who make high ratings on both scales. So there is some evidence of national patterns of consistently high and low ratings for Relative Effort.

The results reported above include all students. We have already noted however that a small number of students report unrealistic levels of effort. In particular, we have noted that 7.2% of responding students report putting more effort

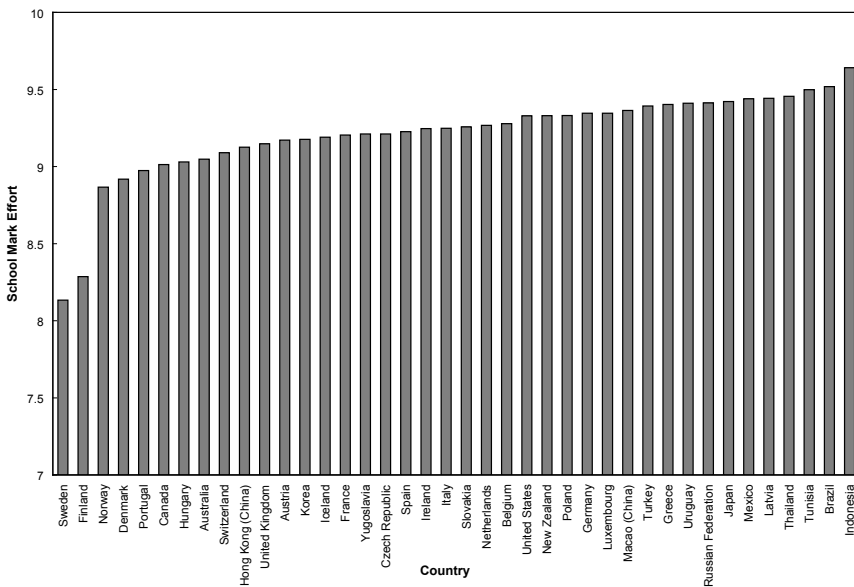


Figure 3. School Mark Effort by country



into PISA than they would into an assessment that counted towards their school mark.

In Figure 5 the proportions of students who gave unrealistic ratings for each country are displayed. Figure 5 shows an atypical pattern for Indonesia—a much higher proportion of *unrealistic raters* when compared to other countries. Further, the four countries with the

highest levels of *unrealistic raters*—Thailand, Tunisia, Brazil and Indonesia—are countries with low mean reading achievement.

A Scandinavian pattern of low numbers of *unrealistic raters* is also evident. Sweden, Finland, Norway and Denmark are in the top five countries showing lowest levels of students reporting unrealistic effort. Overall, countries

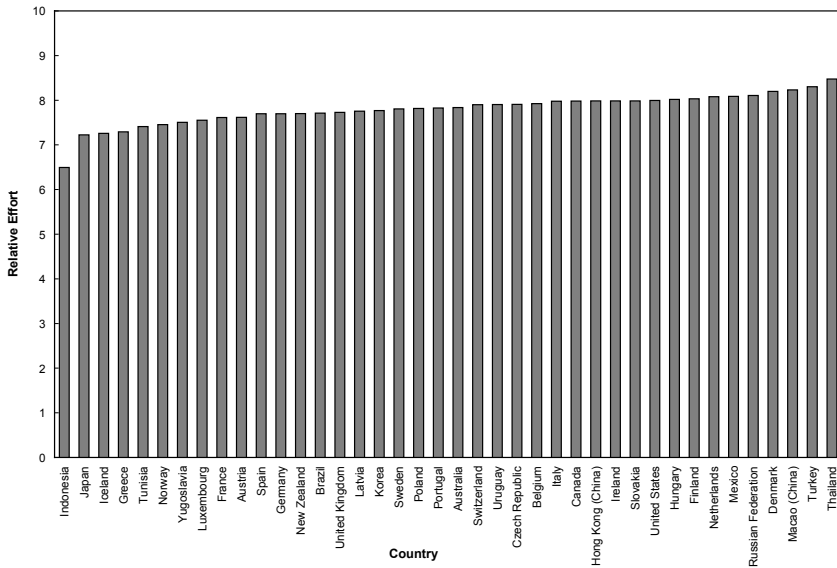


Figure 4. Relative Effort by country

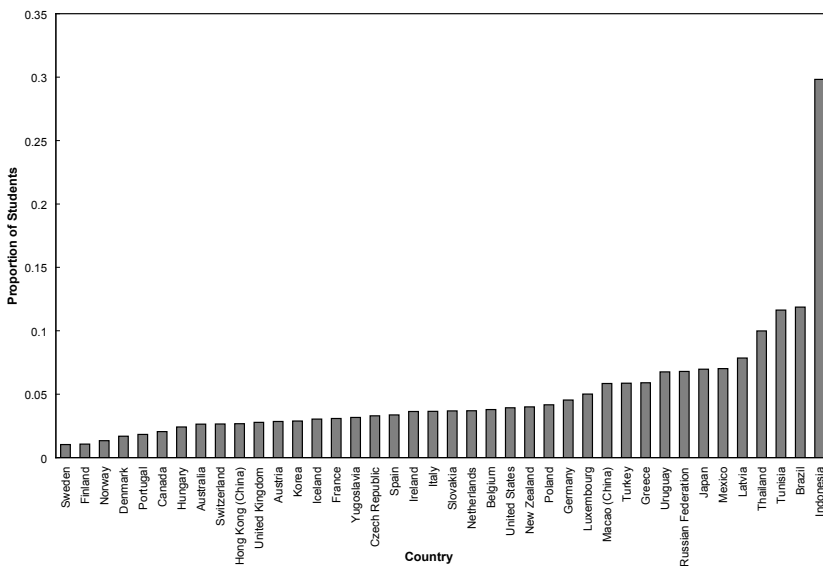


Figure 5. Proportion of students showing unrealistic effort by country

with above average reading compared to the OECD average are reporting low unrealistic effort ratings whilst high levels of unrealistic effort are reported by less developed countries with below OECD averages in reading. It is possible that this pattern is influenced by student ability when reading the different effort scenarios presented in the Effort Thermometer.

Indonesia is an unusual case as 29.8% of students responded unrealistically to the Effort Thermometer. Two hypotheses are offered to explain this pattern. Firstly, unrealistic ratings could be caused by a social desirability bias where participants choose to portray themselves in a favorable way. Secondly, the explanation could be that these students are poor or remedial readers and misunderstood the two sub-scales of PISA Effort and School Mark Effort.

#### *Correlations of types of effort and reading*

Table 3 reports the correlation between country mean achievement and mean effort scores. The results are provided for all countries and for all countries except Indonesia. It seems prudent to examine the behaviour of the data when a country with an atypical effort ratings profile is excluded.

Table 3

#### *Correlations for types of effort and achievement for all countries and excluding Indonesia*

	All countries	All countries except Indonesia
Relative Effort	0.245	0.020
PISA Effort	-0.506	-0.464
School Mark Effort	0.247	0.060
Unrealistic Effort	-0.697	-0.749

Table 3 shows that, at the country level, Relative Effort is positively correlated with reading achievement. If Indonesia is excluded this correlation approaches zero. Overall, the positive correlations for Relative Effort and School Mark Effort disappear if Indonesia is excluded while the negative correlations for PISA Effort and unrealistic raters remain.

Overall, the unrealistic raters are demonstrating lower levels of reading ability and the

proportion of unrealistic raters tends to be greater in countries with lower mean achievement

This relationship could be evidence that these ratings are caused by poor reading skills rather than a social desirability bias. The relationship between unrealistic effort ratings and country level achievement is also displayed in Figure 6.

Figure 6 shows the proportion of unrealistic effort raters by country. The countries are ordered in achievement from Tunisia as the lowest achieving country to Finland the highest achieving country in reading. The peaks in the graph show the anomalies in these types of raters. Again, Indonesia's rating pattern is prominent. The ten highest achieving countries have low levels of students responding unrealistically; that is, less than 0.05% students are making these unrealistic judgements. Also, the Scandinavian pattern of low numbers of unrealistic raters is apparent. Sweden, Finland, Norway and Denmark demonstrate high reading achievement and a low percentage of students making unrealistic ratings.

#### *The influence of unrealistic raters*

The social desirability hypotheses and the remedial reading hypothesis have been proposed to account for the behavior of unrealistic raters. Both hypotheses raise some doubt about the trustworthiness of the Effort Thermometer ratings. Therefore, in subsequent analyses a parallel approach is adopted where analysis for *all students* and for *realistic raters* is undertaken. This approach represents a fuller picture and attempts to distil the issues.

The group *all students* are the students who responded to the Effort Thermometer for PISA 2003. The group *realistic raters* are students from PISA 2003 who responded to the Effort Thermometer in a realistic and sensible fashion.

#### *Investigating some adjustments*

We now further explore the influence that Relative Effort might be having on PISA's headline results: that is, to what extent might differential investment of effort influence the (relative) standings of the countries.

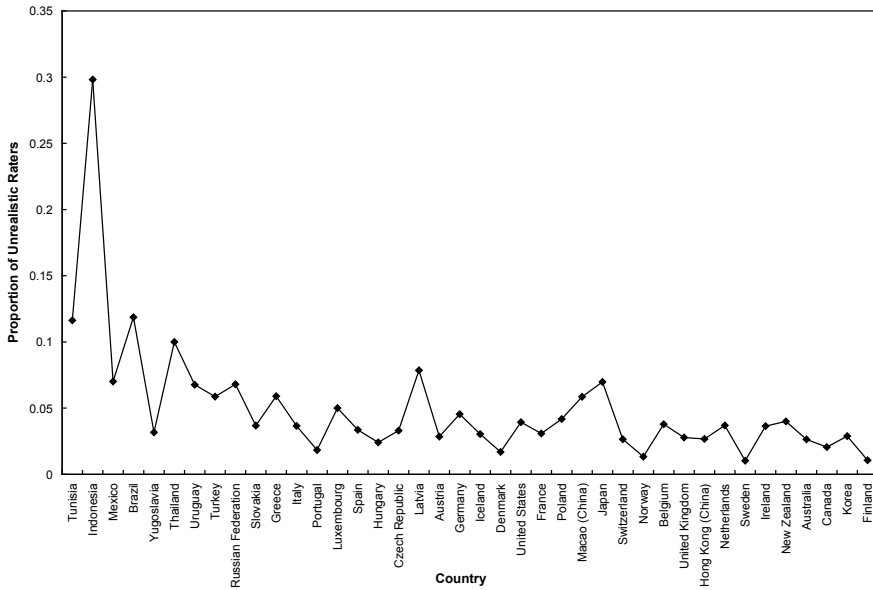


Figure 6. Proportion of students making unrealistic ratings by countries ranked in mean reading achievement order

Table 4 shows the relationship between country and reading performance both for all sampled students and *realistic raters*. The relationship is expressed both as a multiple correlation and an  $R^2$ . These statistics were computed by using students as the level of analysis and regressing reading achievement on a dummy coding of students' country membership.

Table 4 shows that almost 20% of the variation in student performance in reading can be ac-

Table 4

*Multiple R and R<sup>2</sup> when regressing country mean achievement on country using all students and realistic raters*

	All students	Realistic raters
Multiple R	0.444	0.407
R <sup>2</sup>	0.197	0.166

Table 5

*Multiple R and R<sup>2</sup> when regressing country mean achievement on country and Relative Effort using all students and realistic raters*

	All students	Realistic raters
Multiple R	0.449	0.420
R <sup>2</sup>	0.202	0.177

counted for by country. If the analysis is restricted to the *realistic raters* the country variance reduces to 16.6 percent.

The next step is to add effort into the model and examine the influence that Relative Effort has on the between country differences.

Table 5 shows the relationship between country, Relative Effort and reading performance both for *all students* and for the *realistic raters*. The relationship is expressed both as a multiple correlation and an  $R^2$ . These statistics were computed by using students as the level of analysis and regressing reading achievement on a dummy coding of students' country membership and a dummy coding of Relative Effort. Comparing the results with those in Table 4 we note an increase from 19.7% to 20.2% (0.5%) in terms of the variance in student performance that is explained by the model.

Based upon the two models that are summarised in Table 4 and Table 5 it is possible to estimate a raw country mean reading score and a country mean reading score adjusted for Relative Effort. Figure 7 shows a scatter plot of the raw (or unadjusted) means against means that are adjusted for Relative Effort. The squares

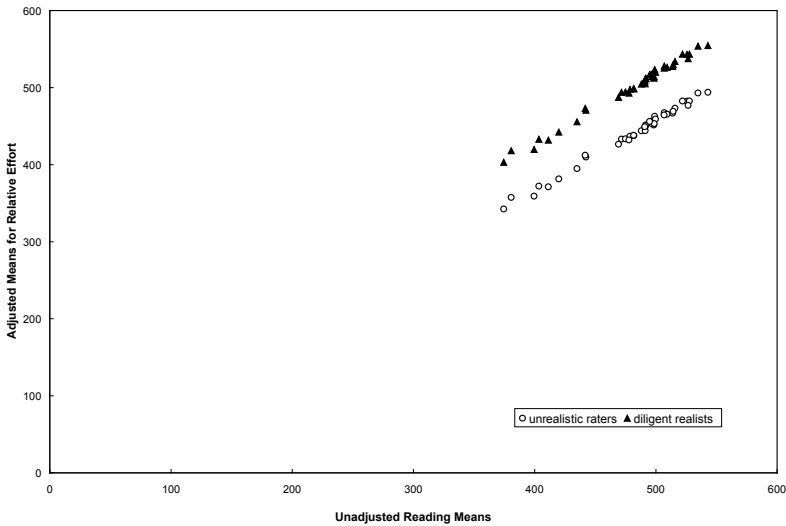


Figure 7. Mean achievement for diligent realists and unrealistic raters

show the predicted means for each country if all students are behaving like *unrealistic raters*. The triangles show the predicted means for each country if all students are behaving like *diligent realists*. The average size of the adjustment is a 46.76 point decline for *unrealistic raters* and a 13.96 point improvement for *diligent realists*.

The mean (across countries) of the estimated raw means is 480.93. The estimated adjusted mean (across countries) of the diligent realists is 500.60. That is, if all students in all countries had behaved like the *diligent realists* then the estimated mean would have been 19.67 points higher.

If analysis is restricted to *realistic raters* then the values are 489.70 for the average raw mean and 500.60 for the mean for *diligent realists*. Therefore, if realistic rating students in all countries were acting like *diligent realists* then the estimated mean for reading achievement would be 11.10 points higher.

The differences vary from 31.27 in Turkey to 11.69 in Finland. The extreme difference is observed for Indonesia where the difference is 37.51 points. However, as discussed earlier, Indonesia presents an atypical case.

The magnitude of and importance of this difference can be contextualized by considering, as examples, each of the following three rela-

tionships that have been observed in PISA data. First, economic, social and cultural characteristics account for one fifth of the student variation in performance in OECD countries (OECD, 2003; Schulz, 2006). For effort we find about 0.5%. The impact of effort, therefore, is substantively small and as shown in Table 4 does not explain differences between countries, as has been argued by some commentators on international studies. Second, the gender difference on the reading scale of retrieving information is a 26 point advantage to females. This difference is just a little larger than is the difference between the performance of all students and *diligent realists*. Third, in OECD countries students from single-parent families have reading scores that are on average 12 points lower than students from other types of families (OECD, 2003). This difference is the same as the difference between the performance of *realistic raters* and *diligent realists*.

In summary it seems that effort effects are not large enough to invalidate the cross-national comparisons. They are of the same order of magnitude as some differences (eg gender differences) that are usually regarded as substantively important.

*Exploring gender, Relative Effort and reading*

This section investigates the gender difference in relation to Relative Effort for students and

for countries. The section also examines the gender gap in reading achievement and then explores the relationship between reading achievement and Relative Effort for males and females. As in the previous sections the practice of investigating Relative Effort will be conducted with parallel analyses using *all students* and *realistic raters*.

A significant goal of schooling is the support of equal educational opportunity for both males and females. Investigation of education achievement has highlighted the fact that this goal is under threat. In reading achievement, there is a growing international concern about the achievement of males. The superiority of females in reading literacy is substantial. However, the gender gap is more pronounced in some countries than others, which suggest gender differences can be ameliorated by educational practices (OECD, 2001).

Results from the PISA 2000 contextual study found that females are more closely engaged in reading. Males tend to read only when required whereas females are more likely to read for enjoyment. Males and females also differ in the types of materials they read voluntarily with males preferring newspapers, comics, e-mails and Web pages whilst females prefer reading fiction (OECD, 2003). Considering these differing patterns in engagement and choice of reading materials it seems prudent to investigate whether test-taking motivation displays a gender bias. If this pattern is evident, it is useful to investigate the effect of this difference in motivation on performance in a reading assessment.

Karmos and Karmos (1984) investigated test-taking motivation through self report scales. They found that girls were significantly more positive in their ratings than the boys but attitudinal scores for boys were more often and more strongly correlated with achievement z-scores. However a study with contradictory results was published by Brown and Walberg (1993). These researchers

found that the motivational effect was the same for boys and girls. Considering the divergent results, further investigation of the effect of gender and its relationship to test-taking motivation and reading achievement seems appropriate.

The results from PISA 2000 display a strong pattern of female advantage in reading achievement. The superior performance of females was reported as being not only universal but large. On average the male and female achievement difference was 32 points or approximately half a PISA proficiency level. This gender divide was typically larger than the difference in mean scores between countries (OECD, 2001).

Results from PISA 2003 reported similar differences. Females showed significantly higher average reading performance than males. The female reading advantage is again reported as half a proficiency level (OECD, 2004a). Based on reading achievement scores for *all students*, the correlation between gender and reading achievement is 0.136. This correlation is equivalent to a difference in mean scores for males and females of 29.41 points on the PISA reading literacy scale.

A continuation of the parallel analysis is adopted in order to investigate whether Relative Effort is different for males and females and whether any such differences, if they do exist, might influence observed performance differences. Table 6 shows calculated means for Relative Effort for *all students* and for *realistic raters*. Standard errors are shown in brackets.

Table 6 shows that female students in both the *all student* and the *realistic raters* groups have higher mean Relative Effort than male students. Additionally the difference in mean Relative Effort is greater for *all students* than it is for *realistic raters*. Because the latter group have responded in a reasonable way to the Effort Thermometer it could be inferred that the value of 0.225 is the

Table 6

*Mean (standard error) Relative Effort for females and males across all countries in PISA 2003*

	Female Relative Effort	Male Relative Effort	Mean difference
All students	7.919 (0.021)	7.614 (0.027)	0.305 (0.029)
Realistic raters	8.482 (0.014)	8.257 (0.016)	0.225 (0.018)

better reflection of male and female difference in effort investment. It should be noted that gender differences for *all students* ( $t = 10.517, p < 0.01$ ) and for *realistic raters* ( $t = 12.5, p < 0.01$ ) are statistically significant but they are small—approximately one quarter of a point on the Effort Thermometer.

A small multiple correlation exists between gender and Relative Effort and reading achievement ( $r = 0.260$ ). Although this correlation is modest, it is double the correlation between gender alone and reading achievement.

We now investigate whether statistical adjustments made for Relative Effort would produce a difference in results regarding average reading achievement and gender at the national level.

To achieve this we fit the following regression model

$$R_i = \alpha_0^* + \alpha_1^* G_i + \beta_1 E_i^1 + \beta_2 E_i^2 + \dots + \beta_{10} E_i^{10} + \varepsilon_i, \tag{1}$$

where  $R_i$  is the reading proficiency of student  $i$ ,  $G_i$  is ‘1’ if student  $i$  is male and ‘0’ otherwise. The ten dummy variables,  $E_i^1, E_i^2, \dots, E_i^{10}$  account for the Relative Effort of student  $i$ . The variable  $E_i^j$  takes the value, ‘1’ if the Relative Effort of student  $i$  is  $j-1$  and it takes the value ‘0’ otherwise. For example if a student is a *realist* then  $E_i^9 = 1$  and all remaining  $E_i^j$  are ‘0’.

Under this model  $\alpha_0^*$  is the estimated mean performance of boys, adjusted for Relative Effort and  $\alpha_1^*$  is the gender difference adjusted for effort. The parameter estimates for this model are shown in Table 7.

Using male *PISA supporters* as the reference group, the poorest performing group in reading is both the male and female *unrealistic raters*. Male unrealistic raters are on average 62.39 points less than the reference group while females are 45.99 points less than the reference group.

Table 7 shows that the highest performing group are the *PISA realists* with Relative Effort score 8. The second highest performing group is the students who had Relative Effort score 7. The

Table 7

*Parameter estimates for the model*

Regression coefficient	Estimate	Standard error
$\alpha_0^*$	472.082	1.937
$\alpha_1^*$	16.399	.988
$\beta_1$	-61.759	2.545
$\beta_2$	-22.653	6.594
$\beta_3$	-5.284	6.539
$\beta_4$	2.009	4.976
$\beta_5$	1.558	4.935
$\beta_6$	7.704	3.265
$\beta_7$	17.444	2.902
$\beta_8$	25.441	2.085
$\beta_9$	27.091	1.912
$\beta_{10}$	25.408	1.967

third highest achieving group is the *diligent realists* who had Relative Effort score 9. The figure shows the poor achievement of the students who made unrealistic ratings. This could be further evidence that either these students were not able to cope with the reading demands of the Effort Thermometer or that poor achievers have much higher levels of compliance than other students.

*Gender and Relative Effort*

Next we consider whether the patterns relating to the gender differences for Relative Effort and observed in the international context are demonstrated at the national level. The global picture reveals that female students have higher mean Relative Effort scores compared to male students. This difference is observed for *all students* and *realistic raters*. Additionally, a modest correlation exists between gender, Relative Effort and reading achievement. This correlation is approximately, by comparison, the size of the correlation observed between socio-economic background and reading achievement (OECD, 2003).

PISA 2003 revealed a consistent pattern of female students outperforming male students in reading. However, the gender difference variation between countries was sizeable, ranging from 58 points in Iceland to 21 points in Korea, Mexico and the Netherlands and 13 points in Macao (China).

The correlations between gender and reading achievement range from negligible—Macao ( $r = 0.090$ ) to noticeable—Iceland ( $r = 0.295$ ) Norway ( $r = 0.229$ ) Austria ( $r = 0.229$ ). Iceland has the biggest gender gap in reading of all countries participating in PISA 2003. Some possible reasons for the gender difference exhibited by the Icelandic results are examined in Olafsson, Halldorsson, and Bjornsson, 2006.

The correlations between gender and reading correspond to differences ranging from 58.34 points to 12.19 points on the PISA reading scale. Countries where the gender difference in reading can be considered large (greater than 47 points) include Iceland, Norway and Austria.

An examination of gender differences for Relative Effort by country allows us to see if higher ratings by females are universal. Figure 8 shows the difference between mean Relative Effort of males and females for *all students*. Figure 9 shows the difference for the *realistic raters*.

Figure 8 shows that female students are reporting higher Relative Effort than male students. Korea and Japan have the smallest difference between males and females while Sweden and Poland have the largest difference.

Figure 9 shows most countries report higher Relative Effort for females. Korea shows higher Relative Effort for males compared to females. However, it should be noted that although the difference between the male and female ratings for these countries is negative, it is also small. Japan and Indonesia have the smallest positive difference between males and females while Poland and Sweden have the largest positive difference. Therefore, limiting the sample to those students who made realistic ratings does have some effect on the rankings of countries in terms of gender differences for Relative Effort. It also reveals the anomalous behaviour of the male students from Korea. However limiting the sample to just *realistic raters* would not necessarily provide the fullest picture of the interplay between motivation and achievement.

*Gender and Relative Effort and reading*

Multiple regressions for gender and Relative Effort and reading were calculated. Some countries show a low correlation: Japan ( $r = 0.199$ ), Korea ( $r = 0.192$ ), Russia ( $r = 0.163$ ) and Tunisia ( $r = 0.169$ ). A modest correlation is shown by Iceland ( $r = 0.385$ ), Latvia ( $r = 0.329$ ), Norway ( $r = 0.364$ ) and Sweden ( $r = 0.327$ ). Adjusting for

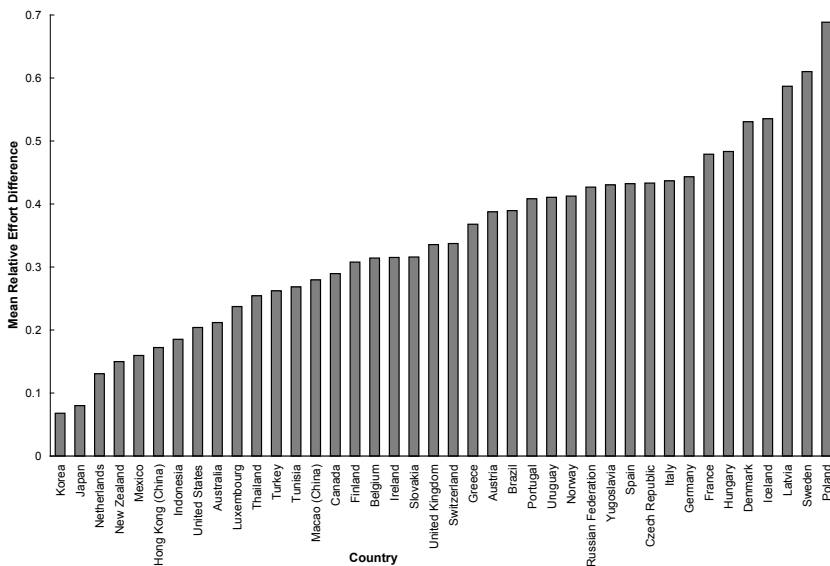


Figure 8. Difference in Relative Effort by gender for all students

gender and Relative Effort gives a clearer picture at the country level of how Relative Effort is influencing the gender gap in reading achievement.

*Adjusting for gender and Relative Effort*

We next investigate whether controlling for gender and Relative Effort has an impact on interpretation of results. We carry out this investigation by running regression model (1) separately for each country.

Figure 10 shows that some countries have negligible differences while other countries have differences that translate into a ten point difference on the reading scale. Countries where adjusting for Relative Effort has a small impact on the gender gap in reading achievement are Indonesia, Slovakia and Canada. Countries where a large impact on the gender gap in reading achievement is observed are Norway, Iceland, Turkey and Belgium.

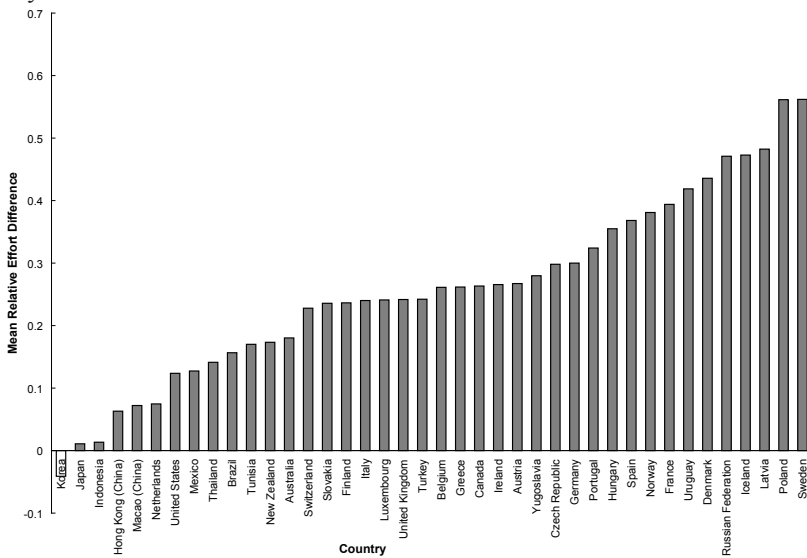


Figure 9. Difference in Relative Effort by gender for *realistic raters*

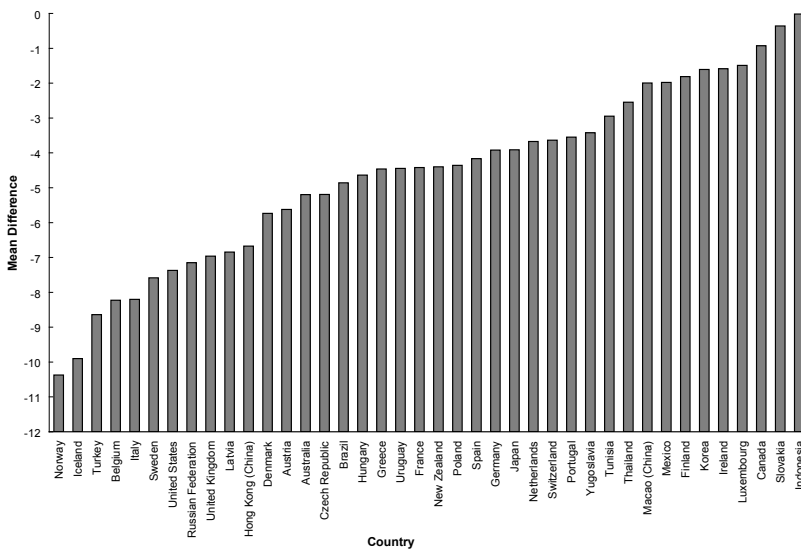


Figure 10. Difference between raw and adjusted means for gender and Relative Effort



In all countries the adjustment for effort reduces the estimate of gender difference. This suggests that part of the observed gender differences in PISA is reflective of a differing level of invested effort by male and female students. In the middle range of countries effort accounts for about five points.

*The Australia and Germany case study*

As described in the first section, Australia, Germany and Norway piloted the Effort Thermometer in PISA 2000, followed by all participating countries in 2003. Australia and Germany were selected as case study countries because they generated Effort Thermometer data from both cycles, which allows us to examine the potential influence that variations in students’ investment in effort may have on the interpretation of trend results for PISA.

An additional reason for the selection of Australia and Germany was the different reaction in these countries to the results from PISA 2000. In Australia the results were disseminated through low key media releases and publications (ACER, 2001). In Germany the results received intense media coverage as the German public was shocked by the finding that their reading mean was below the OECD average (Deutsche Presse Agentur, 2001; Fertig, 2003). The climate between the two cycles also varied for the two countries. In Australia, PISA had little impact and remained relatively unknown. However,

in Germany PISA 2000 results became widely debated (Fertig, 2003).

The data analysed in this section was sourced from three places. The PISA 2003 data all originates from the PISA web site, and was prepared according to PISA international standards. The PISA 2000 achievement data was also sourced from the PISA web site. The PISA 2000 effort data for Australia was sourced from the Australian PISA management centre at the Australian Council for Educational Research and was merged with the achievement data. The PISA 2000 effort data for Germany was sourced from the German PISA management centre at the *Institut für die Pädagogik der Naturwissenschaften* (IPN) and was merged with the achievement data.

Table 8 shows the un-weighted sample sizes for PISA 2000 and PISA 2003 in Australia and Germany. For PISA 2000, 228 schools in Australia and 213 schools in Germany participated; for PISA 2003, 301 schools in Australia and 211 schools in Germany were involved.

According to PISA Technical reports (Adams and Wu, 2002; OECD, 2005) the Australian and German data meet PISA’s strict standards for both PISA 2000 and PISA 2003.

*Effort*

Looking at the distribution of Relative Effort we observe different patterns for Australia and Germany that change over time. Table 9 shows

Table 8  
*Sample characteristics for PISA 2000 and PISA 2003 for Australia and Germany*

	Australia		Germany	
	2000	2003	2000	2003
All students	2806	6335	2438	2315
Males	2629	6216	2574	2299
Females	42	0	61	46
Invalid gender	5477	12551	5073	4660
Total				
	Australia		Germany	
	2000	2003	2000	2003
Realistic raters	2517	5687	2112	2001
Males	2387	5661	2245	2063
Females	16	0	42	39
Invalid gender	4920	11348	4399	4103
Total				

Table 9

*Distribution of students over categories of Relative Effort*

Relative Effort Category	Score	2000		2003	
		Australia	Germany	Australia	Germany
PISA supporters	10	19.3	13.8	16.6	18.4
Diligent realists	9	25.4	18.7	23.5	22.3
PISA realists	8	23.0	22.0	24.7	21.6
	7	12.2	13.0	13.4	11.6
	6	5.2	6.6	5.6	5.7
	5	3.1	4.9	3.5	4.4
	4	1.2	1.7	1.6	1.5
	3	0.7	1.2	0.8	1.1
	2	0.3	0.7	0.6	0.4
PISA cynics	1	0.5	0.6	0.5	0.5
Unrealistic raters	0	3.0	1.9	2.5	4.2
Invalid responses		6.2	14.8	6.8	8.4
Reasonable effort		79.9	67.5	78.2	73.9

the percentage of students for each category of Relative Effort for Germany and Australia for both PISA 2000 and PISA 2003.

The distribution of Relative Effort for Australia and Germany in 2000 indicates that students in Germany show lower expenditure of effort than do Australian students. Comparing Australia over time, we see overall effort declined from 2000 to 2003. There are fewer *PISA supporters* and *diligent realists*. Comparing Germany over time, we see a reverse pattern. There is an increase in *PISA supporters* and *diligent realists* between 2000 and 2003.

Using the combined weighted distribution we can examine what percentage of students is stating that they are investing a reasonable amount of effort in the assessment. We define students who indicate a reasonable level of effort as the *PISA supporters*, the *diligent realists*, the *PISA realists* and the next group of realists (values in italic in Table 9). Overall, for both countries and cycles, 74.9% of students are indicating a conscientious attitude to the assessment in terms of applying reasonable effort. This percentage is a positive indication of the number of students who are trying hard for an assessment that has limited personal consequences.

In contrast, the group labelled the *PISA cynics* is a small and relatively stable group for both countries and both cycles. This group could be a recalcitrant group who are reporting high effort

expenditure for PISA and low effort expenditure for School Mark Effort, which indicates that these students may not be taking PISA seriously. Another noteworthy group is the *unrealistic raters* in Germany in 2003. This group has grown compared to 2000. It is possible that the importance of PISA as a beacon of national achievement could be responsible for the increase.

Table 10 shows the mean Relative Effort for *all students* and *realistic raters* for Australia and Germany. We first note that the mean Relative Effort for Australia is greater than that for Germany in both 2000 and 2003. The difference in 2000 is 0.310 ( $t = 5.08, p < .01$ ), and the difference in 2003 is 0.137 ( $t = 2.69, p < .01$ ). These differences are statistically significant.

Table 10

*Mean (standard error) Relative Effort by country for all students and realistic raters*

Cycle	Country	All students	Realistic raters
2000	Australia	7.948 (.048)	8.208 (.037)
	Germany	7.638 (.038)	7.812 (.033)
2003	Australia	7.836 (.025)	8.049 (.025)
	Germany	7.699 (.044)	8.066 (.034)

While the difference in 2003 is still significant it has reduced, caused by a statistically significant decline in the effort of Australian students (0.112,  $t = 2.08, p < .05$ ) and a small, but non-significant rise for the German students (0.061,  $t = 1.04, p > .01$ ).

The next step in the comparison is to see if the patterns found for *all students* are consistent with the patterns observed for *realistic raters*. In PISA 2000 the *realistic raters* in Germany had a lower mean for Relative Effort than the *realistic raters* in Australia. The difference of 0.396 ( $t = 7.987, p < .01$ ) is statistically significant.

In 2003, the mean for Australia had decreased by a small amount (0.159,  $t = 4.29, p < .01$ ), so the students were now trying a little less hard, whereas the mean for Germany had increased (0.254,  $t = 7.69, p < .01$ ). Students in Germany were now trying harder than students in Australia. However, the difference is not significant (0.017,  $t = .403, p > .01$ ).

The *realistic raters* have higher mean ratings compared to *all students*. Therefore, *realistic raters* in Australia and Germany in 2000 and in 2003 were investing more effort in PISA compared to *all students*. The effort expenditure for Germany increased from 2000 to 2003. The means for the *realistic raters* reveal a larger increase compared to all German students. It could be hypothesized that German students in 2003 were aware of the controversy surrounding the results for 2000 and were prepared to make reasonable attempts for the 2003 assessment.

Table 11 shows Relative Effort means broken down by both PISA cycle and gender. These results show that for *all students* and for *realistic raters* the girls are consistently trying harder than the boys. For *all students* the gender differences in Australia are 0.204 and 0.212 ( $t = 4.285, p < .01$ ) for PISA 2003. In Germany the gender differences are 0.480 ( $t = 7.741, p < .01$ ) and 0.443 ( $t = 6.515, p < .01$ ) respectively. Therefore, for *all students*

the gender differences for both cycles and both countries are statistically significant.

For *realistic raters* the gender differences for Australia by cycle are 0.227 ( $t = 3.771, p < .01$ ) and 0.181 ( $t = 4.568, p < .01$ ). For Germany the gender differences are 0.428 ( $t = 7.824, p < .01$ ) and 0.299 ( $t = 5.067, p < .01$ ). Again, all of these differences are statistically significant. Compared to *all students* the subset of *realistic raters* shows a smaller difference in mean Relative Effort between German boys and girls. In 2000 and in 2003 this difference has decreased and is closer to the Australian difference. In conclusion, it appears that by 2003 the German students as represented by the *realistic raters* are approximating the amount of effort expended by the Australian students.

*Reading*

This sub-section investigates the patterns that the reading results show for the two cycles of PISA, for Germany and Australia, and for male and female students. The reading means for all students are arrayed by cycle, country and gender and are presented in Table 12.

Table 12 shows a non-significant decline in Australia from 2000 to 2003 of 3.407 points ( $t = 0.850, p > .01$ ), while Germany shows an increase from 2000 to 2003 of 7.367 points ( $t = 1.759, p > .01$ ) which is also not significant. So, the results indicate a non-significant change in national reading performance for both countries over time.

Australia outperforms Germany on both occasions, but the difference is substantially reduced in 2003. The performance gap closed from 44.843 points in 2000 to 34.069 points in 2003. The

Table 11

*Mean (standard error) Relative Effort by gender for all students and realistic raters*

Cycle	Country	Gender	All students	Realistic raters
2000	Australia	Female	8.059 (.050)	8.332 (.036)
		Male	7.855 (.066)	8.105 (.055)
	Germany	Female	7.881 (.038)	8.025 (.032)
		Male	7.401 (.059)	7.597 (.051)
2003	Australia	Female	7.943 (.031)	8.140 (.025)
		Male	7.731 (.038)	7.959 (.038)
	Germany	Female	7.929 (.049)	8.220 (.035)
		Male	7.486 (.063)	7.921 (.054)

performance gap between countries for 2000 is significant ( $t = 10.684, p < .01$ ). Although the gap decreases for 2003 it is still significant ( $t = 8.52, p < .01$ ). So, we have significant performance gaps between the two countries.

Girls outperform boys on all four occasions. For PISA 2000 the gender difference was similar in both countries. The gender gap for Australia was 31.574 ( $t = 5.206, p < .01$ ) and for Germany was 34.646 ( $t = 6.929, p < .01$ ). By 2003 the German gender difference had increased to 42.123 ( $t = 7.321, p < .01$ ) compared to 39.339 ( $t = 10.312, p < .01$ ) in Australia. The national gender gaps are significant for both cycles.

The performance of the Australian girls is statistically equivalent across the two cycles, the difference is 1.659 ( $t = 0.318, p > .01$ ). German girls show a slight increase of 10.729, which just fails to reach statistical significance at the 0.05 level ( $t = 1.950, p > .01$ ). For the boys, neither Australia nor Germany shows changes that are statistically significant. The Australian decline is 7.132 ( $t = 1.452, p > .01$ ) and the German increase is 3.252 ( $t = 0.616, p > .01$ ).

In summary, what we have observed is non-significant change in national reading performance over time and non-significant declines and increases for males and females. We have noted the significant performance gaps between the two countries and the significant national gender gaps over cycles. We shall now investigate whether the

same patterns are observed for *all students* and for the *realistic raters*.

Table 13 shows a non-significant decline in achievement for Australia over the two cycles of 3.427 points ( $t = 0.877, p > .01$ ). However, the increase for German students over two cycles is negligible at 0.273 points and not significant ( $t = 0.068, p > .01$ ).

The performance gap narrows from 35.785 points in 2000 to 32.085 points in 2003. The performance gap between countries for 2000 is significant ( $t = 8.602, p < .01$ ). Although the gap decreased for 2003 it is still significant ( $t = 8.620, p < .01$ ). Therefore, the *realistic raters* in Australia still outperform the *realistic raters* in Germany on both occasions. The difference between the two countries is statistically significant and closer for the *realistic raters* than for *all students*.

The superior performance of female students is observed for both countries and for both cycles. In 2000 the gender difference for Australia is 31.574 ( $t = 5.228, p < .01$ ) and for Germany 27.616 ( $t = 6.502, p < .01$ ). By 2003 the German gender difference has increased to 37.558 ( $t = 6.837, p < .01$ ) compared to 36.607 ( $t = 10.085, p < .01$ ) in Australia. The national gender gaps are significant for both cycles.

The performance of the Australian girls is statistically equivalent across the two cycles, the difference is 1.556 ( $t = 0.301, p > .01$ ). German girls show a smaller increase of 5.082 compared

Table 12  
*Mean (standard error) reading scores for all students*

All students		Female	Male	Overall
2000	Australia	547.084 (4.549)	513.218 (4.012)	528.834 (3.397)
	Germany	502.198 (3.875)	467.552 (3.169)	483.991 (2.465)
2003	Australia	545.425 (2.553)	506.086 (2.835)	525.427 (2.126)
	Germany	512.927 (3.905)	470.804 (4.226)	491.358 (3.386)

Table 13  
*Mean (standard error) reading scores for realistic students*

Realistic raters		Female	Male	Overall
2000	Australia	553.337 (4.569)	521.763 (3.949)	536.714 (3.431)
	Germany	515.104 (2.915)	487.489 (3.089)	500.929 (2.353)
2003	Australia	551.781 (2.420)	515.174 (2.706)	533.287 (1.866)
	Germany	520.186 (3.767)	482.628 (3.998)	501.202 (3.221)

to all students, which is not statistically significant ( $t = 1.067, p > .01$ ). For the boys, neither Australian nor German results are statistically significant. The decline for Australia is 6.589 ( $t = 1.376, p > .01$ ) and for Germany 4.861 ( $t = 0.962, p > .01$ ).

With the exception of the performance of German boys, the same patterns of achievement are observed for *realistic raters* as for *all students* in the two cycles. Therefore, we conclude that for *realistic raters* the same patterns observed for *all students* hold. The change in national reading performance over time is not significant. The performance difference between the two countries and the gender gaps over time are significant. The fluctuations of the males and female realistic raters over time are not significant.

In summary, a relationship between reading achievement and Relative Effort has been shown. We have also shown there are differences in Relative Effort between *all students* and *realistic raters*, the two countries, between males and females and over time. In regards to reading achievement we have also shown there are differences between *all students* and *realistic raters*, the two countries, between males and female students in Australia and Germany and in the results from PISA 2000 and 2003.

*Adjusting for Relative Effort*

In this section, we examine the impact that adjusting for Relative Effort has on the interpretation of the performances of students in Germany and Australia. We do so by fitting two alternative regression models to the pooled 2000 and 2003 Australian and German data sets. The first regression model is:

$$R_i = \alpha_0 + \alpha_1 C_i + \alpha_2 Y_i + \alpha_3 G_i + \alpha_4 (C_i \times Y_i) + \alpha_5 (C_i \times G_i) + \alpha_6 (G_i \times Y_i) + \alpha_7 (C_i \times Y_i \times G_i) + \varepsilon_i, \tag{2}$$

where  $R_i$  is the reading proficiency of student  $i$ ,  $C_i$  is '1' if student  $i$  is Australian and '0' otherwise,  $G_i$  is '1' if student  $i$  is female and '0' otherwise,

and  $Y_i$  is '1' if student  $i$  is a year 2003 student and '0' otherwise.

Under this model the parameters can be interpreted as follows:

- $\alpha_0$  is the estimated mean performance of German boys in 2000;
- $\alpha_1$  is the overall country difference;
- $\alpha_2$  is the overall PISA cycle difference;
- $\alpha_3$  is the overall gender difference;
- $\alpha_4$  is the country by PISA cycle interaction;
- $\alpha_5$  is the country by gender interaction;
- $\alpha_6$  is the gender by PISA cycle interaction; and,
- $\alpha_7$  is the three-way interaction between country, PISA cycle and gender.

Model (2) is referred to as the *raw* regression model. The estimates for the parameters for the raw regression model are displayed in Table 14.

Table 14  
*Parameter estimates for raw model*

Regression coefficient	Estimate	Standard error
$\alpha_0$	466.437	3.145
$\alpha_1$	46.177	5.408
$\alpha_2$	4.076	5.535
$\alpha_3$	35.761	5.151
$\alpha_4$	-10.603	7.329
$\alpha_5$	-1.290	7.879
$\alpha_6$	6.653	7.400
$\alpha_7$	-1.785	10.062

A second regression model is given in Equation (3). In this model we add in 10 dummy variables,  $E_i^1, E_i^2, \dots, E_i^{10}$  to account for the relative effort of student  $i$ . The variable  $E_i^j$  takes the value, '1' if the Relative Effort of student  $i$  is  $j-1$  and it takes the value '0' otherwise. For example if a student is a *realist* then  $E_i^9 = 1$  and all remaining  $E_i^j$  are '0':

$$R_i = \alpha_0^* + \alpha_1^* C_i + \alpha_2^* Y_i + \alpha_3^* G_i + \alpha_4^* (C_i \times Y_i) + \alpha_5^* (C_i \times G_i) + \alpha_6^* (G_i \times Y_i) + \alpha_7^* (C_i \times Y_i \times G_i) + \beta_1 E_i^1 + \beta_2 E_i^2 + \dots + \beta_{10} E_i^{10} + \varepsilon_i. \tag{3}$$

Under this model each of the alpha parameters has the same meaning as in model (2) but they are adjusted for the Relative Effort dummy variables. As such the estimates based upon this model are *adjusted* estimates.

The dummy coding of Relative Effort via the  $E_i^j$  variables is such that the *PISA supporters* are the reference category, ie a student who is a *PISA supporter* would have  $E_i^j = 0$  for all  $j$ .

The estimates of the parameters for the adjusted regression model are displayed in Table 15.

Comparing the results obtained from estimating these two regression models allows us to explore the pivotal question—does effort make a difference? After adjusting for Relative Effort we can examine the impact of effort on reading achievement.

Table 16 shows how the parameter estimates from the two regression models can be used to produce estimates of the mean performances of various subgroups of students before and after adjustment for effort. The estimators of the adjusted figures include the parameter  $\beta_{10}$  so that

Table 15

*Parameter estimates for the adjusted model*

Regression coefficient	Estimate	Standard error	Regression coefficient	Estimate	Standard error
$\alpha_0^*$	492.017	3.712	$\beta_1$	-72.742	6.136
$\alpha_1^*$	30.335	4.821	$\beta_2$	-70.341	11.832
$\alpha_2^*$	-6.223	5.059	$\beta_3$	-55.802	11.074
$\alpha_3^*$	24.160	3.850	$\beta_4$	-43.271	10.175
$\alpha_4^*$	0.388	6.564	$\beta_5$	-42.814	7.868
$\alpha_5^*$	4.469	6.666	$\beta_6$	-27.506	5.887
$\alpha_6^*$	10.245	6.167	$\beta_7$	-17.930	4.294
$\alpha_7^*$	-4.287	8.844	$\beta_8$	-7.908	3.579
			$\beta_9$	2.136	2.938
			$\beta_{10}$	13.846	3.608

Table 16

*Estimators of subgroup means based upon the raw and adjusted regression models*

	Raw			
	Australia		Germany	
	2000	2003	2000	2003
Males	$\alpha_0 + \alpha_1$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_4$	$\alpha_0$	$\alpha_0 + \alpha_2$
Females	$\alpha_0 + \alpha_1 + \alpha_3 + \alpha_5$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7$	$\alpha_0 + \alpha_3$	$\alpha_0 + \alpha_2 + \alpha_3 + \alpha_6$
	Adjusted			
	Australia		Germany	
	2000	2003	2000	2003
Males	$\alpha_0^* + \alpha_1^* + \beta_{10}$	$\alpha_0^* + \alpha_1^* + \alpha_2^* + \alpha_4^* + \beta_{10}$	$\alpha_0^* + \beta_{10}$	$\alpha_0^* + \alpha_2^* + \beta_{10}$
Females	$\alpha_0^* + \alpha_1^* + \alpha_3^* + \alpha_5^* + \beta_{10}$	$\alpha_0^* + \alpha_1^* + \alpha_2^* + \alpha_3^* + \alpha_4^* + \alpha_5^* + \alpha_6^* + \alpha_7^* + \beta_{10}$	$\alpha_0^* + \alpha_3^* + \beta_{10}$	$\alpha_0^* + \alpha_2^* + \alpha_3^* + \alpha_6^* + \beta_{10}$

the adjustments are as if Australian and German students were acting like *diligent realists*.

The findings are presented firstly describing gender trends for males then females followed by country trends for Australia then Germany. Again the data will be displayed for *all students* and for the subset of *realistic raters*.

*Gender trends*

Firstly, we look at the performance of boys and whether the trend patterns vary between *all students* and *realistic raters*. The results for the male students from Australia and Germany in 2000 and 2003 are displayed in Table 17.

For male students, Australia shows a decline from 2000 to 2003. This pattern holds for both the raw and adjusted figures for both *all students* and *realistic raters*.

For German male students we see an indication of an increase in performance from 2000 to 2003 in the raw data. However, if this is adjusted for effort we see a small decline. That is, any apparent increase in the performance of males in Germany can be explained by their increased level of effort.

The mean for the male realistic students show a decline in reading from 2000 to 2003 of 4.23 points. When adjusted for effort the decline is still evident but has increased to 6.26 points.

So adjustment reveals a larger decline in reading achievement than the raw figures indicate.

In conclusion, for boys in Germany reading achievement has not improved after adjustment for expenditure of effort. This finding suggests that the improvement is not reflecting better reading outcomes but a better, more positive attitude to taking the test. It is possible that the interpretation of the improved results for boys in PISA 2003 did not account for increased positive motivation and results would have been interpreted as higher reading proficiency. Equivalently the improvement could be attributed do an underestimate of male performance in 2000. Considering the impact that PISA 2003 results had in Germany the apparent improvement of male students in reading would be a welcomed but erroneous interpretation.

The results for the female students from Australia and Germany in 2000 and 2003 are displayed in Table 18.

In Australia the performance for all females declines slightly from 2000 to 2003. The difference is small (1.659). This holds also after adjustment (0.123). In Germany there is an improvement of 10.729 points from 2000 to 2003. After adjustment this is reduced to 4.022 points. It can be noted that some, but not all of the increase in female scores in Germany can be accounted for by effort.

Table 17

*Raw and adjusted reading means for all males and for male realistic raters*

All students		2000	2003	Realistic raters		2000	2003
Australia	Raw	512.614	506.087	Australia	Raw	521.778	515.174
	Adjusted	536.198	530.363		Adjusted	536.956	531.138
Germany	Raw	466.437	470.513	Germany	Raw	486.600	482.375
	Adjusted	505.863	499.640		Adjusted	505.473	499.215

Table 18

*Raw and adjusted reading means for all females and for female realistic raters*

All students		2000	2003	Realistic raters		2000	2003
Australia	Raw	547.085	545.426	Australia	Raw	553.337	551.781
	Adjusted	564.827	564.950		Adjusted	565.710	565.483
Germany	Raw	502.198	512.927	Germany	Raw	515.104	520.186
	Adjusted	530.023	534.045		Adjusted	530.018	533.874

The picture for the female realistic raters shows that for Australia from 2000 to 2003 there is a small decline (1.556) which is negated with adjustment (0.227). For Germany reading performance improves by 5.082 points from 2000 to 2003 but with adjustment this increase is made smaller (3.856).

For both countries, the adjusted scores are higher than their matching raw figures because *diligent realists* have been used as the adjustment. However, using a *diligent realist* reference does not change the findings.

The gender trends for the two countries are now examined. The results for the Australian male and female students from 2000 and 2003 are displayed in Table 19.

For Australia a widening in the gender difference is seen over time. The female advantage in reading is increasing. Further, the change seems to be due to a decline in the performance of male students rather than any change in the performance of the female students. While all means are increased by an adjustment for Relative Effort, the pattern remains the same.

This picture of a widening of the gender gap is present for *realistic raters* but is not as pronounced as for *all students*. Therefore, in the Australian setting by considering *all students* and *realistic raters*, it appears that effort expenditure is not influencing the reading results from PISA.

The results for the German male and female students from 2000 and 2003 are displayed in Table 20.

The pattern is more interesting in Germany. Here the raw figures show an increase in both the performances of male and female students. The increase appears marginally larger for females than males. After adjustment, the improvement by the female students is reduced and the change in the performance of the male students becomes negative. The adjusted results seem more consistent with the adjusted results from Australia.

The picture for *realistic raters* is revealing. It shows that German males did not improve their reading achievement from 2000 to 2003 and this pattern is stronger in the adjusted case.

In summary effort explains some of the increase in female performance while disguising a decline in male performance. The adjusted results show a widening of gender differences that is more consistent with Australia. However, when controlling for investment of effort the results point to a decline and not an increase in reading achievement.

### Conclusion

Expenditure of effort on PISA can be captured and measured to enable comparison across cycles and countries. Effort represented by the variable Relative Effort is constructed with scores labelled to reflect differing levels of

Table 19

*Raw and adjusted reading means for all Australian students and for Australian realistic raters*

All students		2000	2003	Realistic raters		2000	2003
Females	Raw	547.085	545.426	Females	Raw	553.337	551.781
	Adjusted	564.827	564.950		Adjusted	565.710	565.483
Males	Raw	512.614	506.087	Males	Raw	521.778	515.174
	Adjusted	536.198	530.363		Adjusted	536.956	531.138

Table 20

*Raw and adjusted reading means for all German students and for German realistic raters*

All students		2000	2003	Realistic raters		2000	2003
Females	Raw	502.198	512.927	Females	Raw	515.104	520.186
	Adjusted	530.023	534.045		Adjusted	530.018	533.874
Males	Raw	466.437	470.513	Males	Raw	486.600	482.375
	Adjusted	505.863	499.640		Adjusted	505.473	499.215



effort expenditure. Reassuringly, most students make judgements relating to the expenditure of effort on PISA that reflects a realistic assessment of the situation. Generally students, according to their reports expend equal or less effort on PISA compared to a situation where their PISA results would contribute to their school marks. A small percentage of students make judgements that are unrealistic. Generally these students exhibit lower reading ability.

The main finding that the expenditure of effort is fairly stable across a majority of countries may be instrumental in countering the claim that differential effort invalidates international comparisons. Effort is related to reading achievement with an effect size similar to variables such as single parent family structure, gender and socio-economic background. Countries reporting higher levels of unrealistic effort, which may be due to a social desirability bias, also demonstrate lower levels of reading achievement.

The relationship between Relative Effort, reading achievement and gender shows that girls report higher levels of effort in undertaking PISA than boys. The correlation between reading achievement and gender doubles with the inclusion of effort and controlling for effort decreases the difference in reading achievement between boys and girls.

The main findings on the Australian-German case study using data from PISA 2000 and 2003 show that the expenditure of effort is less in Germany than in Australia but Germany has improved its effort expenditure from 2000 to 2003. Based on the performance of the students who made realistic effort judgements, the reading achievement difference between the two countries was less in 2003 compared to 2000. Lastly, the influence of effort may be responsible for the improved reading performance of German boys in 2003.

### References

ACER. (2001). *PISA in brief from Australia's perspective*. Camberwell, NSW, Australia: Author.

Adams, R. J., and Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.

Baumert, J., and Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441-462.

Bracey, G. (1999). The ninth Bracey report on the condition of public education. *Phi Delta Kappan, 81*(2), 147-168.

Brown, S. M., and Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136.

Deutsche Presse Agentur (2001, 4 December). *Miserable Noten für deutsche Schüler*. Frankfurter Allgemeine Zeitung.

Fertig, M. (2003). *Who's to blame? The determinants of German students' achievement in the PISA 2000 study*. Discussion Paper No. 739, Institute for the Study of Labor (IZA) retrieved March, 2003, from <http://www.iza.org>

Holliday, W. G., and Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Education Forum, 67*(Spring), 250-257.

Karmos, A. H., and Karmos, J. S. (1984). Attitudes towards standardised achievement tests and their relation to achievement test performance. *Measurement and Evaluation in Counseling and Development, 17*, 56-66.

King, G., Murray, C., Salomon, J., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(1), 191-207.

Kiplinger, V. L., and Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment, 3*(2), 111-133.

- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2002). *German scale handbook for PISA 2000*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.
- Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis*, 4, 419-437.
- OECD. (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: Author.
- OECD. (2003). *Literacy skills for the world of tomorrow. Further results from PISA 2000*. Paris: Author.
- OECD. (2004a). *First results from PISA 2003. Executive Summary*. Paris: Author.
- OECD. (2004b). *Learning for tomorrow's world. Results for PISA 2003*. Paris: Author.
- OECD. (2005). *Technical report for the OECD programme for international student assessment 2003*. Paris: Author.
- Olafsson, R. F., Halldorsson, A. M., and Bjornsson, J. K. (2006). Gender and the urban-rural differences in mathematics and reading: An overview of PISA 2003 results in Iceland. In J. Mejdning and A. Roe (Eds.), *Northern lights on PISA 2003: A reflection from the Nordic countries* (pp. 185-198). Copenhagen, Denmark: Nordic Council of Ministers.
- O'Neil, H. F., Jr., Sugrue, B., and Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3, 135-157.
- O'Neill, H. F., Sugrue, B., Abedi, J., Baker, E. L., and Golon, S., (1997). *Final report of experimental studies on motivation and NAEP performance* (CSE Technical Report 427). Los Angeles, CA. University of California: CREST.
- O'Neill, H. F., Abedi, J., Lee, C., Myoshi, J., and Mastergeorge, A. (2004). *Monetary incentives for low-stakes tests* (CSE Technical Report 625). Los Angeles, CA. University of California: CREST.
- Pintrich, P. R., and Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2<sup>nd</sup> ed.). Upper Saddle River, NJ: Merrill Prentice Hall.
- Schulz, W. (2006, April). *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and 2003*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Turner, R., and Adams, R. J. (2007). The Programme for international student assessment: An overview. *Journal of Applied Measurement*, 8, 237-248.
- Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, 30, 1-21.
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.
- Wise, S. L., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17.
- Wolf, L. F., and Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.