International Education Journal, 2006, 7(2), 127-149.
ISSN 1443-1475 © 2006 *Shannon Research Press*.
http://iej.com.au

Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view

Petra Lietz
International University Bremen, Germany *p.lietz@iu-bremen.de*

Results of a previous meta-analysis of gender differences in reading achievement at the secondary school level (Lietz, in press) showed significant differences between major assessment programs. Thus, the gender gap in favour of girls was more pronounced for the assessment programs conducted by the National Assessment of Educational Programs in the United States (NAEP), for the more recent assessment programs in Australia and the Programme for the International Student Assessment (PISA) conducted by the OECD. In contrast, no such effect was found for earlier studies conducted by the International Association for the Evaluation of Educational Achievement (IEA), namely the International Reading Comprehension Study 1970-71 and the International Reading Literacy Study 1990-91.

Hence, this article seeks to investigate whether or not an effect exists that could be associated with the time period in which a study was conducted. In other words, the article examines whether or not the reasons for the greater gender differences in more recent assessment programs might be related to the scaling of reading scores before and after 1992.

Reading achievement; scaling of scores; meta-analysis;
hierarchical linear modelling; gender differences

INTRODUCTION

The research reported in this article extends across two major areas, one content-related area, namely gender differences in reading, and one method-related area, namely meta-analysis. Each of these areas is discussed briefly below.

Gender Differences in Reading Achievement

The view prevails that boys perform better than girls in mathematics (Aiken and West 1991, Johnston and Dunne 1996, Husen 1967, Keeves 1988, Tracy 1987) and the natural sciences whereas the reverse holds in reading, social studies and languages (Dedze 1995, Plisko 2003, Thorndike 1973, Wagemaker et al. 1996). A closer examination of the research on reading, however, reveals that the matter is not as clear-cut as it might appear and that results can be grouped into two main categories: one showing evidence of girl's superiority over boys in reading achievement, and one providing little or no evidence of gender differences. Thus it can be argued that the research provides some support for the existence of a gender gap in reading performance in favour of female students, while some studies and reviews dispute this finding. However, these studies provide inconclusive evidence with regard to the extent of gender differences in reading at the secondary school level. Hence, a more systematic approach to integrating research findings, namely statistical meta-analysis (Glass, McGaw and Smith 1981, Hunter and Schmidt 2004) is suggested and discussed in greater detail below.

Meta-Analysis

For 40 years and more, reports of research findings concerned with the magnitude of the difference between two means have recorded the size of an effect in terms of a standardised difference. This standardised difference was first referred to as an 'effect size' by Cohen (1969). The effect size was calculated by dividing the difference between the means of the two independent groups, by the pooled standard deviation of the two groups. Moreover, Cohen showed how it was related to the point biserial correlation coefficient, not only by multiplying the correlation coefficient by 2, when two large groups were of approximately equal size, but also by using another multiplying factor for unequal sized groups.

Subsequently, the term 'meta-analysis' involving an analysis of effect sizes was introduced by Glass (1976, 1977) to denote a systematic integration of research findings on a specific topic and has been developed further as an analytical technique (Rosenthal 1984, Hedges and Olkin 1985). The need for a more systematic way of integrating prior research than narrative research reviews was introduced as a reaction to criticisms aimed at the social sciences by funding agencies and the public as to whether or not any progress was being made in terms of establishing some statements of knowledge from the seemingly abounding and contradictory evidence generated from many research projects in the social sciences (Light and Smith 1971).

As Hunter and Schmidt (2004, p. 16) emphasised: "In many areas of research, the need today is not for additional empirical data but for some means of making sense of the vast amounts of data that have been accumulated." Moreover, they point out that the narrative integration of research findings has serious shortcomings in that this strategy of integrating research results often leads to different conclusions if done by different people. Statistical meta-analysis, in contrast, as a quantitative way of integrating research findings should lead to the same conclusion, regardless of the person applying the procedure.

Thus, the challenge in the social sciences, in general, and in educational research in particular, is to integrate systematically and quantitatively findings from the large number of research studies that have been undertaken in order to contribute empirically verified facts to the cumulative body of knowledge.

None of the meta-analyses undertaken to date have focused specifically on gender differences in reading. In addition, advances in hierarchical linear modelling (HLM) have occurred that allow for the clustered nature of meta-analytic data to be taken into account more appropriately. Thus, Raudenbush and Bryk (2002) argued that the main purpose of a meta-analysis was to examine the extent to which effects reported in the results of primary studies were consistent and to disentangle what part of the variance in study results was due to sampling error and what component was due to actual treatment implementation. As a consequence, Raudenbush and Bryk (2002) proposed an empirical Bayes meta-analysis as a special application of the two-level hierarchical linear model. In this model, the outcome variable, namely the effect sizes from the different studies, was allowed to vary randomly at the first level while, at the second level, study characteristics were used to explain possible differences in the outcome variable. In other words, the Level-1 analyses were aimed at investigating the extent of the variability in effects sizes of primary studies, while at Level-2 possible sources of this variation might be examined. This extension to two levels was based on the use of ordinary regression models in research synthesis proposed originally by Hedges and Olkin (1983).

In summary, meta-analysis is a systematic way to synthesise findings of research studies on a certain topic. After a systematic search and retrieval of relevant studies, the results are scaled to a common unit of measurement, expressed as effect sizes, usually *d* (Cohen 1988) and allowance is made for different sources of error, in particular, sampling error. The assumption of the meta-analytic approach is that these disattenuated effect sizes are all estimates of a common effect that underlies a whole population of studies. Where variation in effect sizes emerges that is not due to sampling error, the analysis seeks to explain those differences in terms of variation arising from the different contexts and characteristics of the primary studies. As a result of this process, meta-analysis allows the: (a) estimation of effect size parameters, (b) explanation of differences in estimates of effect size, (c) examination of stronger estimates of effect sizes in particular situations, and (d) modelling of factors producing effects in different contexts and under different conditions.

Method

It has been argued (e.g. Cook et al. 1992) that meta-analyses frequently suffered from a lack of transparency with regard to the inclusion or exclusion of primary studies. In order to increase transparency, a summary of the principles guiding the selection of primary studies whose results entered the current meta-analysis is given in Table 1.

Authors have differed in their views on which primary studies to include in a meta-analysis. Slavin (1984, 1986), for example, argued that only primary studies of sound methodological quality should be included in a meta-analysis. Glass et al. (1981), on the other hand, claimed that the breadth of the available evidence should be used when synthesising the current state of knowledge in a particular research area. This view was also supported by Kulik and Kulik (1989) who argued that meta-analyses with a high quality approach to selecting primary studies were often left with too few studies to allow the statistical analysis of the results.

It should be noted that over and above the criteria given in Table 1, no further evaluation of studies was undertaken to determine the inclusion or exclusion of studies entering the current meta-analysis.

In Appendix 1 an overview of the studies included in this meta-analysis is provided whereby national studies or authors analysing data from national studies are listed first, followed by international assessment programs. After the sequential study number in Column 1, the name of the study or the name of the author who reported the study is listed in the second column and followed by information about the country in which the study was conducted in the third column.

The data that are used in the meta-analysis are provided in Columns 4 to 8. The first of these columns contains the effect size in the form of Cohen's *d*. Effect size (ES) is defined by Cohen (1988, p. 8) as follows:

> …it is convenient to use the phrase "effect size" to mean "the degree to which the phenomenon is present in the population", or "the degree to which the null hypothesis is false". Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero.

The reason for Cohen's emphasis on effect sizes stemmed from his criticism of the widespread use of significance tests. Cohen pointed out that the reliance on such tests was misleading not only in that a number of assumptions underlying these tests were frequently not met but in that these tests also provided less information than was possible. While a significance test provided information only as to whether or not the null hypothesis was false, the effect size provided additional information regarding the specific degree to which the hypothesis was false.

Table 1: What the meta-analysis is (not) about

Qualifier	Not about	About
English, verbal ability	Studies that used Grade in English or only general verbal ability as a measure.	Studies had to include some measure of reading comprehension or reading achievement in the language of instruction.
Academic achievement	Studies that did not separate out different aspects of academic achievement – and for example combined mathematics and reading in a single outcome measure were not included.	Studies had to include some measure of reading comprehension or reading achievement in the language of instruction.
Language	Not reading as part of foreign language learning.	Focus was on mother-tongue reading or reading in the languages of instruction.
Information provided	Policy papers, discussion/opinion papers, narrative reviews.	Studies had to provide some data amenable to meta-analysis (means, correlations, regression/path coefficients).
Level of schooling	Primary school level.	Secondary school level (i.e. Grade 6 or 12-year-old students to Grade 12 or 18- year-old students).
Type of variable	Studies that used reading as a predictor, mediator or moderator.	Studies that used reading achievement or reading comprehension as the outcome variable or which focused on correlating various factors or variables with reading achievement.
Reading dimension	Comprehension of a specific type of text or using reading for a specific purpose (e.g. RL's 'documents', 'expository', 'narrative' domains or PISA's 'retrieving', 'interpreting' and 'reflecting' and 'evaluation' skills).	An overall score of performance in reading.
Type of student	Samples that focused on students with disabilities, ethnic minority students.	Samples that were representative of mainstream secondary school students.
Level of data collection	If teacher ratings of student achievement were used; analyses reported at school level (e.g. headmaster studies).	Studies had to focus on student-level variables. Information provided by students.
Type of information	If results were not separated in studies of primary and secondary school students.	Information on effect sizes (e.g. correlation coefficient or mean differences) had to be reported for secondary school students).
Type of publication	Dissertations.	Journal articles (as retrieved from a search using 'secondary' and 'student factors' and 'reading achievement' or 'reading performance' in Eric, Web of Science and PsycINFO and selected according to the criteria in this table) or published study reports.
Date of study	Prior to 1970 or after 2002.	1970-2002

> Thus whether measured in one unit or another, whether expressed as a difference between two population parameters or the departure of a population parameter from a constant or in any other suitable way, the ES can itself be treated as a parameter which takes the value zero when the null hypothesis is true and some other specific nonzero value when the null hypothesis is false, and in this way the ES serves as an index of degree of departure from the null hypothesis. (Cohen, 1988, p. 10)

The way in which to interpret the effect size of Cohen's *d* is as follows. If *d* is calculated to be 0.2, then the means differ by two-tenths of a standard deviation. According to Cohen (1988, p.21) *d* is a pure number, which is freed of dependence upon any specific unit of measurement. A value of 2.0 for *d* indicates that the means differ by two standard deviations. An examination of the effect sizes in the third column of Appendix 1 reveals that values range from –0.87 (Study 57), indicating higher achievement of male students, through 0.00 (Studies 106, 143, 144), indicating no gender differences in reading achievement, to 0.59 (Study 86), indicating a higher performance by female students by about six-tenths of a standard deviation.

The column that follows the effect size is labelled '*v*' which is the squared standard error of *d* (Raudenbush and Bryk 1985; for further details on how '*v*' was calculated, see Equations 2 and 3 below). In the next column, a '1' is assigned if the reading test was administered in English to the whole or the majority of the sample and a '0' if the test was administered in a language other than English. Through the inclusion of this variable in the analysis, it is intended to investigate the potential impact of whether or not the test is administered in English on the variation in gender differences in reading. This is particularly interesting for those assessment programs in which test design takes place in English while tests are administered in many different languages (i.e. PISA, RC, RL). In Column 7, information regarding the mean age of the sample for each study is recorded in order to examine whether or not the possible gender gap in reading increases or decreases with age.

The next column is labelled 'time' and indicates whether a study was undertaken prior to or after 1991. Thus, results from the Reading Literacy Study were assigned a '0' as it was conducted in 1990-91 whereas data provided by the PISA-2000 assessment were assigned a '1' as they had been collected in and after 1992. The reason for choosing 1991 as a cut-off point was the fact that it was only after that date that many testing programs started to use procedures for eliminating at least in part the effects of measurement error from the estimated scores (see Adams, 2005; Wu, 2005) as well as using plausible values in their reports and analyses. Thus, this dummy variable was generated to allow for the examination of possible effects stemming from the way in which reading scores were calculated.

COMMENT ON PARTICULAR MAJOR STUDIES

Below, a short description is given of the assessment programs from which most of the primary study results in the meta-analysis are taken, including information regarding the way in which reading scores were calculated in each program.

Reading Comprehension Study

The first large-scale cross-national survey of reading was conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 15 education systems as the Reading Comprehension Study which formed part of IEA's Six Subject Survey in 1970-71. The reading comprehension test consisted of eight passages and 52 multiple-choice test items that were designed to measure four categories, namely the ability to: (a) follow the organisation of a passage; (b) respond to questions that were specifically answered in the passage; (c) draw inferences from a passage; and (d) identify the writer's purpose. Items were administered to a representative sample of 14-year-old students in each of the participating education systems (Thorndike 1973). In all analyses, Thorndike (1973) used test scores corrected for guessing as indicators of reading performance. These were also the scores used in the current meta-analysis.

Reading Literacy Study

The Reading Literacy Study was the next study of reading performance conducted by IEA in 1990-91. This time, 31 education systems participated at the 14-year-old level (Population B). As in the first study, samples representative of the target population were drawn in each country under the supervision of an international sampling referee. The design of the reading test had shifted from an emphasis on skills to an emphasis on different types of reading materials, namely narrative, expository and documents. As a consequence, students had to answer a total of 89 multiple-choice items relating to 19 passages (Elley 1994). Reading scores based on the one-parameter model developed by Rasch (1960) were calculated as indicators of performance in reading, whereby one overall reading score was calculated as well as three separate ones, one for each domain. While most of the reporting was undertaken by domain, the score used in the current meta-analysis is the overall score for male and female students from Population B for each country that participated in the study (Elley, 1994, p.106).

Programme for International Student Assessment (PISA)

In the late twentieth century, the Organisation for Economic Co-operation and Development (OECD) launched its Programme for International Student Assessment (PISA) with the main aim to compare the performance of students towards the end of compulsory schooling in key subject areas, namely Mathematics, Reading and Science across its member countries. The focus of the first round of data collection in 2000, in which a total of 43 OECD and non-OECD member countries participated, was on reading. The reading test assessed performance on five processes, namely: (a) retrieving information, (b) forming a broad general understanding, (c) developing an interpretation, (d) reflecting on and evaluating the content of a text, and (e) reflecting on and evaluating the form of a text. Items were of the multiple choice as well as the open constructed-response type and related to continuous and non-continuous texts. Each participating country had to survey a nationally representative sample of 15-year-old students and comply with the sampling guidelines of the OECD (Adams and Wu 2002).

In PISA-2000, two types of reading scores were calculated, namely Warm's (1985) weighted likelihood estimator (WLE) and Bayesian estimation procedures with plausible values (PV) (Adams and Wu 2002). While the weighted likelihood estimator uses the actual score a student obtained as the most likely, plausible values are random numbers that are…

> […] drawn from a distribution of scores that could be reasonably assigned to each individual-that is, the marginal posterior distribution. As such, plausible values contain random error variance components and are not optimal as scores for individuals. Plausible values as a set are better suited to describing the performance of the population. (Adams and Wu 2002, p. 107)

For the international PISA-2000 data set, six WLEs were calculated for each student, one for each of the subject areas tested, namely mathematics, reading and science and three for the reading sub-scales, namely retrieving information, interpreting texts and reflection and evaluation. In addition, 30 plausible values were generated for each student: five for each of the three subject areas and five for the three reading sub-scales. The country-level average scores used in this meta-analysis were the first plausible value mean score (PV1read) for male and female students for the overall reading scale, weighted by the population student weight (w_fstuwt)[1].

NAEP Studies

The National Assessment of Educational Progress (NAEP) is an assessment program run by the National Center for Education Statistics (NCES) in the United States Department of Education.

[1] The PISA 2000 technical report (Adams and Wu, 2002) recommends the application of the student weight (w_fstuwt) for all between country-analyses such as the application in this meta-analysis. The report also recommends that ideally, analyses should be repeated for each of the five plausible value estimates. This was not done in the current analysis which used the first plausible value (PV1Read) only. To illustrate how close the population estimates for plausible values are, an example is given from the German PISA 2000 data set.
For girls (all weighted by student population weight): PV1Read=501.9074; PV2Read=502.2901; PV3Read=502.2903; PV4Read=502.4483; PV5Read=502.0534.
For boys (all weighted by student population weight): PV1Read=467.7509; PV2Read=468.7154; PV3Read=467.0083; PV4Read=467.9008; PV5Read=466.3843.

Since 1969, NAEP has conducted studies in a number of subject areas, including reading, to assess achievement levels of nationally representative student samples in Grades 4, 8, and 12. In the most recent reading test design, students were assessed on four aspects of reading. These covered the: (a) forming a general understanding; (b) developing interpretation; (c) relating information in the text to own knowledge and experience; and (d) examining content and structure, which required critical evaluation and an appreciation of the effects of text features such as irony, humour and organisation. To this end, the reading comprehension test employed multiple-choice questions, designed to test students' understanding of individual texts, as well as their ability to integrate and synthesise ideas across the texts and constructed-response questions, which required students to construct their own answers (Plisko 2003).

Over the more than 35 years that NAEP has been the so-called 'Nation's report card' in the United States, the way in which reading scores were calculated has changed as NAEP has used Bayesian estimation procedures and plausible values for its more recent assessment programs (see Beaton 1987; Campbell et al. 2000; Gorman 2005). Thus, the data employed in the current meta-analysis from NAEP assessments between 1971 and 1980 used scores corrected for guessing while the assessments between 1992 and 2003 used plausible values and weighted likelihood estimates.

Australian Studies

In Australia, data on the reading performance of secondary school students were available from a number of studies. They included the 1975 and 1980 studies Australian Studies in School Performance (ASSP) and Australian Studies in Student Performance (1980), the Youth in Transition Study (YIT) in 1989, and the Longitudinal Surveys of Australian Youth (LSAY) that were conducted in 1995 and 1998. The ASSP data included national samples at both ages 10 and 14 years, whereas the Youth in Transition Study and the longitudinal surveys collected data from 14-year-olds only (Rothman 2002).

The reading tests used in these various studies were not the same. The 1975 test was designed to assess minimum competency, and therefore focused on the lower levels of achievement, while the later tests generally covered a wider range of student performance. However, all tests contained a number of common items, which were used in the analysis of trends in reading achievement over time (Marks and Ainley 1996).

The Monitoring Standards in Education (MSE) program in Western Australia started with the Random Sample assessment program in 1990 with data collections that occurred in 1992, 1995, 1997, 1999 and 2001 whereby ten per cent of students in each of Grades 3, 7, and 10 were tested. In 1998, the Western Australian Literacy and Numeracy Assessment (WALNA) population testing began with Grade 3 students. Subsequently, the assessment of Grade 5 was introduced and the Grade 7 was also included. Data collection from Grade 10 students has continued to be undertaken as part of the Random Sample assessment program. Reading performance was assessed on a range of texts that included continuous texts, for example poems, media releases, narrative extracts, as well as non-continuous texts such as charts or tables.

COMMENT ON STATISTICAL PROCEDURES EMPLOYED

It might be argued that the focus of the current meta-analysis on gender differences in reading achievement at the secondary school level was sufficiently narrow to allow for a relatively straight-forward investigation. Unfortunately, this was not the case. Studies that were retrieved as a result of the literature search differed markedly not only in design, sample size, scope and the scale of the reading score but also in the reporting of results. Thus, results were frequently not reported in terms of standardised effect sizes but in terms of correlation coefficients, regression coefficients from single-level and multi-level analyses, sums of squares, percentage differences or mean differences. Hence, some form of standardisation of the results reported by the different studies was required in order to arrive at a metric-free effect size (ES) that could be processed further in the meta-analysis. The formulae that were employed in the conversion of correlation coefficients, standardised scores, and proportions of test items answered correctly to standardised effect sizes are given in Appendix 2.

As the next step, a so-called 'v-known' hierarchical linear model analysis (Raudenbush et al. 2001, Hox 1995) was undertaken. V-known models may be considered a special case of a two-level hierarchical linear model. In general, hierarchical linear models seek to take into consideration the nested structure of many data sets whereby, for example, students (Level-1) are nested within schools (Level-2). In these instances, variation in the outcome variable at Level-1, frequently a measure of student performance in some subject area, is sought to be explained by variables at Level-1, for example, Gender or Socio-economic status or Homework effort as well as by variables at Level-2, for example, School resources, Size of school, or Location of school. In a meta-analysis the hierarchical structure of the data is such that the within-study variation is modelled at Level-1 while between-study variation is used at Level-2 to explain variability at Level-1. In other words, multilevel modelling as applied to meta-analysis proceeds in two steps. First, it examines whether the within-study results at Level-1 are homogeneous or heterogeneous. If results are homogeneous, the effect sizes may be combined into one average outcome. If the results are heterogeneous, between-study characteristics such as Type of study design or Type of study participants are examined at Level-2 to see whether or not they contribute to explaining differences in results. The reason why Raudenbush and Bryk (2002) labelled these multilevel models for meta-analysis 'v-known models' stems from the fact that the variability at Level-1 is considered to be sampling variability which is known if the relevant sampling distribution and sample sizes are known. Below, the v-known HLM meta-analysis is worked through for the current meta-analysis based on the considerations put forward by Raudenbush and Bryk (2002, p. 208-210).

The effect size (ES) estimate, d_j, for most of the studies listed in Appendix 1 is the standardised mean difference between the average reading scores for female and male students:

$$d_j = \left(\overline{Y}_{Ej} - \overline{Y}_{Cj}\right)/ S_j \qquad [1]$$

where

$\overline{Y}_{Ej}$ is the average reading score for the experimental group, that is, female students;

$\overline{Y}_{Cj}$ is the average reading score for the control group, that is, male students;

S_j is the pooled, within-group standard deviation.

Each of the effect sizes recorded in Appendix 1 is an estimate of the population mean difference between the experimental group, which in this context, consists of female students and the control group, which, in this instance, is male students. Thus, in the second study in the Appendix 1, for example, female students score one-tenth of a standard deviation higher than male students.

With reference to Hedges (1981), Raudenbush and Bryk (2002) stated that d_j follows a normal distribution with variance V_j where

$$V_j = (n_{Ej} + n_{Cj})/(n_{Ej}n_{Cj}) + \delta_j^2 /[2(n_{Ej} + n_{Ej})] \qquad [2]$$

and assert that "it is common to substitute d_j for δ_j and then assume that V_j is "known"" (Raudenbush and Bryk, 2002, p. 209). While the above formula applies to instances where effect sizes are calculated on the basis of mean differences, the following formula applies to effect sizes calculated on the basis of correlation coefficients:

$$V_j = 1/(n_j - 3) \qquad [3]$$

In order to define the hierarchical model for meta-analytic problems, equations have to be formulated at two levels. The model at Level-1, that is the within-study model, is:

$$d_{ij} = \delta_j + e_{ij} \qquad [4]$$

where each of the effect sizes for the 147 studies in the current meta-analysis is considered to be one estimate of the underlying population parameter δ_j plus the sampling error associated with each estimate, e_{ij} with $e_{ij} \sim N(0, V_j)$ (where *i* is the within study subsample, and *j* is the study sample).

At Level-2, study characteristics and random error are considered to predict the unknown effect size δ_j. Thus, the model at Level-2, that is the between-study model, is:

$$\delta_{\mathrm{j}} = \gamma_0 + \gamma_1 \mathrm{W}_{1\mathrm{j}} + \gamma_2 \mathrm{W}_{2\mathrm{j}} + \gamma_3 \mathrm{W}_{3\mathrm{j}} + \mathrm{u}_{\mathrm{j}} \qquad [5]$$

Where: $W_{1j},\ldots,W_{3j} =$ are the study characteristics, namely:

(a) Two general predictor variables:

$W_1 =$ English as the language of test administration, $W_2 =$ Age.

(b) Whether a study was conducted up to and including 1991 or from 1992 onwards:

$W_3 =$ Time,

$\gamma_0,\ldots,\gamma_3$ are the regression coefficients associated with the study characteristics W_1 to W_3,

u_j is Level-2 random error where $u_j \sim N(0, \tau)$.

In order to combine the two-levels into a single model, δ_j in Equation 4 has to be replaced by δ_j from Equation 5:

$$d_{\mathrm{ij}} = \gamma_0 + \gamma_1 \mathrm{W}_{1\mathrm{j}} + \gamma_2 \mathrm{W}_{2\mathrm{j}} + \gamma_3 \mathrm{W}_{\mathrm{sj}} + \mathrm{u}_{\mathrm{j}} + \mathrm{e}_{\mathrm{ij}} \qquad [6]$$

In summary, the Level-1 outcome variable in the meta-analysis is the effect size which quantifies the difference between male and female students' performance in reading reported by each study. In case the variation in effect sizes is found not to be due to chance, the analysis reveals the extent to which variables W_1 to W_3 contribute to explaining the variance.

W_1 and W_2 are specified to examine the potential effects of two variables, namely whether or not English is the language of testing and the average age of the students in a particular study. This allows the examination of two questions. First, since most of the instrument construction for international tests is undertaken in English and with an interest in gender equitable materials in that language, gender differences may be less pronounced in countries where English is the language of instruction and test administration. Second, as male students mature later than female students and reading is basically a process of reasoning (Lietz 1996; Thorndike 1917), gender differences may decrease with increasing age.

The effect sizes used in this meta-analysis were taken from large-scale national and international studies. Thus, in order to examine possible systematic impact on effect sizes of the way in which scaled performance scores were calculated from 1992 onwards, the dummy variable W_3 (Time) was created to indicate whether a study was undertaken up to and including 1991 (dummy code '0') or from 1992 onwards (dummy code '1'). In this way, it was possible to investigate whether or not any systematic difference, associated with the time period in which a study was conducted, emerged. Evidence supporting the introduction of such a time variable is given in the section below entitled 'Some problems involved in comparing effect sizes from different testing programs'.

RESULTS

As noted above, the first step in a meta-analysis using HLM was to examine whether the effect sizes from the different primary studies are homogenous or heterogeneous. In the case where heterogeneity could be ascertained, an analysis was undertaken to investigate the way in which possible study characteristics could contribute to the variability in effect sizes.

Testing the Null Model

It can be seen in Table 2 that the estimated grand-mean effect size, the intercept in the model, is positive and small, $\gamma_{10}(G10)$=0.18, which means that, on average, female secondary students performed about 0.18 standard deviation units above male secondary students. It should be noted that the number of degrees of freedom is 146, one fewer than the number of studies in the analysis, as one degree of freedom is needed for the estimation of the unconditional model. The only parameter to be estimated is the intercept.

Furthermore, the estimated variance of the effect parameter is 0.024 with a standard deviation of 0.15 indicating important variability in the effect sizes. Moreover, the Chi-square value (2557.46) and corresponding p-value (0.000) confirm that this variance is not due to chance and that the residual variance is significantly different from zero. As a consequence, the analysis can proceed to examine which of the predictor variables that reflect study characteristics are able to explain this variance.

Table 2: Final estimation of fixed effects: Unconditional 'v-known' model

```
----------------------------------------------------------------------
Standard Approx.
Fixed Effect       Coefficient    Error    T-ratio    d.f. P-value
----------------------------------------------------------------------
For EFFSIZE, B1
INTRCPT2, G10        0.184245    0.014021   13.141      146     0.000
----------------------------------------------------------------------
Final estimation of variance components:
----------------------------------------------------------------------
Random Effect      Standard     Variance     df    Chi-square   P-value
                   Deviation    Component
----------------------------------------------------------------------
EFFSIZE, U1        0.15420       0.02378     146   2557.46428    0.000
----------------------------------------------------------------------
Statistics for current covariance components model
-------------------------------------------------
Deviance = 964.115997       df = 2
```

Some Problems Involved in Comparing Effect Sizes from Different Testing Programs

While all the testing programs under consideration in this article are concerned with gender differences in achievement, the present information is associated with comparisons that are obtained in many different ways, which are discussed in some detail in Appendix 2. Moreover, the testing programs most probably calculated their sampling variance estimates in different ways in attempting to take into consideration the hierarchical structure of the sample data. Efforts made to develop a set of procedures for this study in order to achieve uniformity in the calculation of effect sizes and estimates of '*v*' proved to be not only frustrating but also unrewarding. Consequently, it was assumed in an earlier analysis (Lietz, in press), that, in general, a particular testing program would use a common procedure across the different studies over time, and allowance could be made for a treatment effect for each of the different testing programs by including dummy variables indicating whether a study belonged to the Reading Comprehension, the Reading Literacy, the PISA, the NEAP or the Australian Testing Program. Likewise, two dummy variables were included in the analysis to indicate the two main bases for estimating effect sizes, namely means and correlations and the corresponding procedure to estimate '*v*'. In Table 3 the results of this earlier analysis (Lietz, in press) are recorded. These analyses included the aforementioned dummy variables plus whether or not English was the language of testing and age as Level-2 predictors. In the table, regression coefficients, their standard errors, t-values associated with each of these predictors as well as the approximate degrees of freedom and p-values that were obtained in initial analyses of the data (Lietz, in press) are presented.

Results showed only small difference between the effect sizes calculated from means (ESMEAN G18 = 0.160) and effect sizes calculated from correlations (ESCORR G19 = 0.188). In contrast, differences between the estimates of the effect sizes for the Reading Comprehension Study (RC G13 = -0.076), the Reading Literacy Study (RL G14 = 0.017) and PISA (PISA G15=0.235) were substantially large. This evidence suggested that the way in which the variance estimates employed in the different methods of estimating effects sizes warranted closer attention.

It was the substantial differences between the coefficients for the different testing programs shown in Table 3 that led to a re-examination of the effect size data. In particular, it became interesting to examine whether the differences may not be so much stemming from the different testing programs per se but be a consequence of different procedures for calculating test scores that were introduced in the early 1990s.

Table 3: Final estimation of fixed effects: 'v-known' model with all predictors included

```
-------------------------------------------------------------------------------
                                   Standard              Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f.  P-value
-------------------------------------------------------------------------------
 For       EFFSIZE, B1
   INTRCPT2, G10          -0.129615   0.162295    -0.799    137      0.425
    ENGLISH, G11           0.021762   0.029647     0.734    137      0.463
        AGE, G12           0.000490   0.009474     0.052    137      0.959
         RC, G13          -0.076360   0.059346    -1.287    137      0.198
         RL, G14           0.016580   0.040229     0.412    137      0.680
       PISA, G15           0.235441   0.038756     6.075    137      0.000
       NEAP, G16           0.181077   0.047319     3.827    137      0.000
         OZ, G17           0.206835   0.039553     5.229    137      0.000
     ESMEAN, G18           0.159932   0.059703     2.679    137      0.008
     ESCORR, G19           0.187569   0.079254     2.367    137      0.018
-------------------------------------------------------------------------------
 Final estimation of variance components:
-------------------------------------------------------------------------------
 Random Effect          Standard      Variance     df     Chi-square  P-value
                        Deviation     Component
-------------------------------------------------------------------------------
  EFFSIZE,       U1       0.09866       0.00973   137      1050.40170    0.000
-------------------------------------------------------------------------------
 Statistics for current covariance components model
--------------------------------------------------
 Deviance  = 882.930946            df = 2
```

In order to summarise the problems raised in this section, it is recognised that in this article the author is attempting to bring together in a meta-analysis the results obtained from the calculation of effect sizes and estimates of '*v*' using very different and perhaps in certain cases possibly inappropriate procedures. Results of the earlier analysis presented in Table 3 showed that these different procedures could possibly be allowed for through the use of dummy variables for the different testing programs. Nevertheless, what is clear is that the comments being made in informal discussions about changes in gender differences in levels of reading performance, being due to changes in reading habits between boys and girls, and the effects of watching TV or working at computers are not warranted until more work is undertaken to examine the procedures used in the different studies that have been undertaken over time. Hence, the following section reports results of a meta-analysis which includes time as a predictor at Level 2.

Change in Recorded Effect Sizes Over Time

In order to examine the potential effect of time on the extent of gender differences, a HLM model which includes the predictors specified in Equations 5 and 6 above was examined and the results are presented in Table 4. Note that the degrees of freedom are now reduced to 143 as, in addition to the intercept, three potential Level-2 predictors, namely Age, whether or not English was the language of testing and Time, needed to be estimated.

Of the three possible predictors only one emerges with a significant effect whereas the remaining two do not contribute to explaining the variability in effects sizes. Thus, Age and whether or not English (ENG) was the language of test-administration do not emerge as significant predictors of gender differences. In other words, the gender gap does not decrease with age, which may have supported the maturational viewpoint whereby reading comprehension is also a function of maturity and, since boys mature at a later age, differences between boys' and girls' reading performance may decrease with increasing age. Likewise, there is no evidence to suggest that gender differences are more or less pronounced in countries where English is not the language of test administration.

However, the impact of the variable Time on the effect size is positive $\gamma_{13}G(13) = 0.24$ and highly significant (*p=0.00*). The way in which this variable is coded means that studies prior and up to 1991 receive the lower ('0') code while studies from 1992 onwards are assigned the higher ('1') code. As a consequence, because the effect of this variable is estimated to involve a gender difference in favour of girls of about 0.24 units higher for studies that had been conducted since 1992 than for those studies that were undertaken prior to that year.

Table 4: Final estimation of fixed effects: 'v-known' model with predictors included

```
--------------------------------------------------------------------------------
                                   Standard              Approx.
    Fixed Effect         Coefficient   Error       T-ratio   d.f.  P-value
--------------------------------------------------------------------------------
 For        EFFSIZE, B1
    INTRCPT2, G10           0.100440   0.113192     0.887     143     0.375
         ENG, G11          -0.017951   0.018880    -0.951     143     0.342
         AGE, G12          -0.001950   0.007544    -0.258     143     0.796
        TIME, G13           0.243677   0.018519    13.159     143     0.000
--------------------------------------------------------------------------------
 Final estimation of variance components:
--------------------------------------------------------------------------------
 Random Effect           Standard      Variance      df     Chi-square  P-value
                         Deviation     Component
--------------------------------------------------------------------------------
  EFFSIZE,       U1        0.08786       0.00772    143      984.49931    0.000
--------------------------------------------------------------------------------

 Statistics for current covariance components model
 ----------------------------------------------
 Deviance  = 776.796393            df = 2
```

In order to arrive at the final hierarchical model, the two between-study variables that did not contribute significantly to explaining differences in effect sizes, namely English and Age were removed from the model. Results of the final model in which only the variable Time is included as a predictor are shown in Table 5.

The intercept in Table 5 is positive and small (0.06), and not significantly different from zero. This finding, in addition to the contrasting results for the intercepts presented in Tables 2, 3 and 4, provides evidence that given the data in this analysis male students performed at a slightly lower level in reading than did female students. However, a sizable and significant positive effect is recorded for Time $\gamma_{11}G(11) = 0.25$ which indicates that since 1992 girls outperformed boys to a considerably greater extent when compared with studies up to and including 1991.

A comparison of the deviance values allows an evaluation of the three models under review, namely the unconditional model, the model which includes all three predictors and the final model with only time as a predictor. Thus, the deviance which is highest for the unconditional model with a value of 964.1 is reduced to 776.8 for the second model. For the final model, in turn, the deviance is further reduced to a value of 743.2 which indicates that the last model provides the best fit to the data, and the removal of the non-significant variables of Age and English yield a better fitting model to the data.

Table 5: Final estimation of fixed effects: 'v-known' model with 'Time' as a predictor

```
--------------------------------------------------------------------------------
                                       Standard             Approx.
    Fixed Effect         Coefficient   Error      T-ratio   d.f. P-value
--------------------------------------------------------------------------------
 For        EFFSIZE, B1
    INTRCPT2, G10           0.059692   0.013339     4.475    145     0.000
        TIME, G11           0.247168   0.018041    13.700    145     0.000
--------------------------------------------------------------------------------

 Final estimation of variance components:
--------------------------------------------------------------------------------
 Random Effect           Standard      Variance    df   Chi-square   P-value
                         Deviation     Component
--------------------------------------------------------------------------------
  EFFSIZE,       U1        0.08751       0.00766   145    1043.16574 0.000
--------------------------------------------------------------------------------

 Statistics for current covariance components model
 --------------------------------------------------
 Deviance  = 743.151878         df = 2
```

The variance estimates of the unconditional model (0.02378) and the final model (0.00766) can be used to calculate the proportion of variance explained in study results. Thus, the final v-known model explains 67.8 per cent ((0.02378-0.00766)/0.02378) of the variance in the data. Complementary information is provided by the chi-square (1043.17) and p-values (0.000) computed for the estimated variance of the effect parameters in the final model of 0.007 which corresponds to a standard deviation of 0.087 and indicates that important variability still exists in the effect sizes. Thus, while the between-study variable Time included in the final v-known model explains about two-thirds of the differences in effect sizes a moderate amount of variability remains to be explained by factors other than those included in this analysis.

CONCLUSIONS

In this study, a meta-analysis of large-scale studies between 1970 and 2002 in the area of reading achievement at the secondary school level with a focus on gender differences was conducted. The meta-analysis was conceptualised as a special application of a two-level hierarchical linear model whereby in a first step, it was examined whether the effect sizes differed more than could be expected due to sampling error. Once results had been ascertained to be sufficiently heterogeneous, characteristics at Level-2 were examined and the way in which they could explain differences between effect sizes at Level-1. Level-2 variables included in the hierarchical linear model covered the age of study participants, and whether or not a study was conducted in a country where English was the language of test administration. In addition, because of the results from an initial meta-analysis which suggested that gender effects were more pronounced in more recent assessment programs a variable indicating whether studies had been conducted prior to or after 1992, was introduced into the analyses.

It is seen that (a) gender differences exist across the 147 studies under review that are not due to chance; and (b) about two-thirds of the variance associated with these differences can be explained by the introduction of a Time variable into the meta-analysis.

Thus, the gender gap in favour of girls is even more pronounced for the assessment programs that have been conducted since 1992. Possible explanations for the origins of these greater differences could be related to item selection procedures or contextual changes surrounding reading in society. Such explanations would appear unlikely, given the stringent psychometric procedures to investigate item bias, in particular with respect to Gender, that had been employed in the large reading assessment programs under review. Likewise, there was little evidence of a general decline in societal support for reading aimed particularly at boys since 1992. Thus, it might be a reasonable explanation that the increase in gender differences for more recent assessment programs might stem from changes in the way in which performance were calculated prior to and after 1992. More specifically, the change to using Bayesian estimation procedures and plausible values or weighted likelihood estimates might have introduced some systematic bias into the effect size indexes as a consequence of a reduction in the within group variance. Alternatively, it might be argued that either prior to 1992 or after 1992 the estimates made of gender differences in reading achievement were basically wrong, because inappropriate estimates of between group variance were being employed in the calculation of effect sizes. Consequently, any discussion of change over time in gender differences in reading achievement and possibly other aspects of educational performance would be inappropriate until the issues raised in this article are resolved.

APPENDIX 1:
STUDIES IN THE META-ANALYSIS
IN ALPHABETICAL ORDER OF AUTHOR OR STUDY

No.	Study/Author	Country	ES(*d*)	*v*	English	Mean Age	Time
1	ASSP 1975	Australia	0.090	0.001	1	14.00	0
2	ASSP 1980	Australia	0.110	0.001	1	14.00	0
3	YIT 1989	Australia	0.080	0.001	1	14.00	0
4	LSAY 1995	Australia	0.190	0.000	1	14.00	1
5	LSAY 1998	Australia	0.230	0.000	1	14.00	1
6	WA monitoring 1992	Australia	0.313	0.003	1	12.00	1
7		Australia	0.344	0.003	1	15.00	1
8	WA monitoring 1995	Australia	0.344	0.003	1	12.00	1
9		Australia	0.389	0.003	1	15.00	1
10	WA monitoring 1997	Australia	0.193	0.003	1	12.00	1
11		Australia	0.448	0.003	1	15.00	1
12	WA monitoring 1999	Australia	0.406	0.003	1	12.00	1
13		Australia	0.434	0.003	1	15.00	1
14	WA monitoring 2001	Australia	0.306	0.000	1	12.00	1
15		Australia	0.496	0.004	1	15.00	1
16	WA monitoring 2002	Australia	0.230	0.000	1	12.00	1
17	Fuller et al. 1994	Botswana	0.143	0.000	1	15.00	1
18		Botswana	0.192	0.000	1	16.00	1
19	GambellandHunter2000	Canada	0.237	0.042	1	13.00	1

20		Canada	0.247	0.042	1	16.00	1
21	Glossop et al. 1979	England	-0.155	0.006	1	15.00	0
22	Gorman et al. 1982	Engl., Wales, Nth. Ireland	0.013	0.001	1	15.75	0
23		Nth. England	0.014	0.004	1	15.00	0
24		Midlands	-0.040	0.005	1	15.00	0
25		Sth. England	0.025	0.003	1	15.00	0
26		Wales	-0.023	0.005	1	15.00	0
27		Nth. Ireland	0.136	0.004	1	15.00	0
28	Youngman 1980	UK	0.040	0.003	1	12.00	0
29		UK	0.283	0.003	1	12.00	0
30	Hogrebe et al 1985	USA,HSB-80	-0.050	0.000	1	17.00	0
31		USA,HSB-80	-0.090	0.000	1	15.00	0
32	LevineandOrnstein 1983	USA,NAEP-71	0.056	0.005	1	13.00	0
33		USA,NAEP-71	0.048	0.005	1	17.00	0
34		USA,NAEP-75	0.056	0.005	1	13.00	0
35		USA, NAEP-75	0.040	0.005	1	17.00	0
36		USA, NAEP-80	0.048	0.005	1	13.00	0
37		USA, NAEP-80	0.038	0.005	1	17.00	0
38	NAEP 2003	USA	0.220	0.005	1	13.00	1
39	NAEP 2002	USA	0.180	0.005	1	13.00	1
40	NAEP 1998	USA	0.280	0.005	1	13.00	1
41	NAEP 1994	USA	0.300	0.005	1	13.00	1
42	NAEP 1992	USA	0.260	0.005	1	13.00	1
43	NAEP 2002	USA	0.320	0.005	1	17.00	1
44	NAEP 1998	USA	0.320	0.005	1	17.00	1
45	NAEP 1994	USA	0.280	0.005	1	17.00	1
46	NAEP 1992	USA	0.200	0.005	1	17.00	1
47	NeumanandProwda 1982	Connecticut 1978-79, USA	0.120	0.000	1	13.00	0
48		Connecticut 1978-79, USA	0.100	0.000	1	16.00	0
49	HedgesandNowell1995	USA, NELS-88	0.090	0.005	1	13.00	0
50		USA, NLS-72	0.050	0.005	1	17.00	0
51		USA, NLSY-80	0.180	0.005	1	18.50	0
52	OaklandandStern1989	Texas, USA	0.006	0.003	0	10.50	0
53	Project Talent 1960	USA	0.150	0.005	1	15.00	0
54	ShillingandLynch 1985	Pennsylvania, USA	0.161	0.005	1	13.00	0
55	Johnson 1973-74	Canada	0.172	0.041	1	12.00	0
56		England	-0.250	0.039	1	12.00	0
57		Nigeria	-0.870	0.038	1	13.00	0
58		USA	0.103	0.041	1	12.00	0
59	PISA2000	Australia	0.330	0.001	1	15.00	1
60	PISA2000	Austria	0.250	0.001	0	15.00	1
61	PISA2000	Belgium	0.330	0.001	0	15.00	1
62	PISA2000	Canada	0.320	0.000	1	15.00	1
63	PISA2000	Czech Republic	0.370	0.001	0	15.00	1
64	PISA2000	Denmark	0.250	0.001	0	15.00	1
65	PISA2000	Finland	0.510	0.001	0	15.00	1
66	PISA2000	France	0.290	0.001	0	15.00	1
67	PISA2000	Germany	0.340	0.000	0	15.00	1
68	PISA2000	Greece	0.370	0.003	0	15.00	1
69	PISA2000	Hungary	0.310	0.002	0	15.00	1
70	PISA2000	Iceland	0.400	0.000	0	15.00	1
71	PISA2000	Ireland	0.290	0.001	1	15.00	1
72	PISA2000	Italy	0.380	0.001	0	15.00	1
73	PISA2000	Japan	0.300	0.004	0	15.00	1
74	PISA2000	Korea	0.140	0.001	0	15.00	1
75	PISA2000	Luxembourg	0.270	0.000	0	15.00	1
76	PISA2000	Mexico	0.210	0.001	0	15.00	1

77	PISA2000	New Zealand	0.460	0.001	1	15.00	1
78	PISA2000	Norway	0.430	0.001	0	15.00	1
79	PISA2000	Poland	0.360	0.002	0	15.00	1
80	PISA2000	Portugal	0.240	0.002	0	15.00	1
81	PISA2000	Spain	0.240	0.001	0	15.00	1
82	PISA2000	Sweden	0.370	0.001	0	15.00	1
83	PISA2000	Switzerland	0.300	0.002	0	15.00	1
84	PISA2000	UK	0.250	0.001	1	15.00	1
85	PISA2000	US	0.280	0.005	1	15.00	1
86	PISA2000	Albania	0.590	0.005	0	15.00	1
87	PISA2000	Argentina	0.440	0.005	0	15.00	1
88	PISA2000	Brazil	0.160	0.001	0	15.00	1
89	PISA2000	Bulgaria	0.480	0.005	0	15.00	1
90	PISA2000	Chile	0.250	0.005	0	15.00	1
91	PISA2000	Hong Kong	0.150	0.005	1	15.00	1
92	PISA2000	Indonesia	0.200	0.005	1	15.00	1
93	PISA2000	Israel	0.150	0.005	1	15.00	1
94	PISA2000	Latvia	0.530	0.005	0	15.00	1
95	PISA2000	Liechtenstein	0.320	0.002	0	15.00	1
96	PISA2000	Macedonia	0.510	0.005	0	15.00	1
97	PISA2000	Peru	0.060	0.005	0	15.00	1
98	PISA2000	Romania	0.130	0.005	0	15.00	1
99	PISA2000	Russia	0.380	0.002	0	15.00	1
100	PISA2000	Thailand	0.420	0.005	1	15.00	1
101	PISA2000	Netherlands	0.300	0.001	0	15.00	1
102	RC 1970-71	Belgium(Fl.)	0.100	0.036	0	14.00	1
103	RC	Belgium(Fr.)	0.345	0.056	0	14.00	0
104	RC	Chile	-0.242	0.010	0	14.00	0
105	RC	England	0.201	0.007	1	14.00	0
106	RC	Finland	0.000	0.014	0	14.00	0
107	RC	Hungary	0.040	0.005	0	14.00	0
108	RC	India	0.040	0.007	1	14.00	0
109	RC	Iran	-0.060	0.033	0	14.00	0
110	RC	Israel	-0.060	0.008	1	14.00	0
111	RC	Italy	0.040	0.003	0	14.00	0
112	RC	Netherlands	-0.060	0.021	0	14.00	0
113	RC	New Zealand	0.040	0.014	1	14.00	0
114	RC	Scotland	-0.140	0.015	1	14.00	0
115	RC	Sweden	0.120	0.011	0	14.00	0
116	RC	USA	0.080	0.007	1	14.00	0
117	RL1990-91	Trin and Tobago	0.299	0.011	1	14.40	0
118	RL	Thailand	0.304	0.007	1	15.20	0
119	RL	Ireland	0.284	0.007	1	14.50	0
120	RL	Canada(BC)	0.259	0.005	1	13.90	0
121	RL	Sweden	0.188	0.007	0	14.80	0
122	RL	Finland	0.215	0.015	0	14.70	0
123	RL	Hungary	0.192	0.007	0	14.10	0
124	RL	United States	0.153	0.006	1	15.00	0
125	RL	Iceland	0.167	0.007	0	14.80	0
126	RL	Italy	0.123	0.006	0	14.10	0
127	RL	Netherlands	0.118	0.006	0	14.30	0
128	RL	Cyprus	0.110	0.020	0	14.80	0
129	RL	Germany(E)	0.096	0.010	0	14.40	0
130	RL	Belgium(Fr.)	0.077	0.007	0	14.30	0
131	RL	Botswana	0.140	0.007	1	14.70	0
132	RL	Hong Kong	0.078	0.006	1	15.20	0
133	RL	New Zealand	0.054	0.008	1	15.00	0

134	RL	Philippines	0.077	0.004	1	14.50	0
135	RL	Slovenia	0.079	0.007	0	14.70	0
136	RL	Denmark	0.052	0.005	0	14.80	0
137	RL	Germany(W)	0.051	0.005	0	14.60	0
138	RL	Norway	0.056	0.007	0	14.80	0
139	RL	Spain	0.062	0.003	0	14.20	0
140	RL	Switzerland	0.041	0.003	0	14.90	0
141	RL	Venezuela	0.033	0.006	0	15.50	0
142	RL	Greece	0.015	0.007	0	14.40	0
143	RL	Nigeria	0.000	0.013	1	15.30	0
144	RL	Singapore	0.000	0.007	1	14.40	0
145	RL	France	-0.059	0.008	0	15.40	0
146	RL	Portugal	-0.133	0.008	0	15.60	0
147	RL	Zimbabwe	-0.283	0.007	1	15.50	0

APPENDIX 2:
CALCULATION OF EFFECT SIZES FOR STUDIES IN THE META-ANALYSIS

1. For the Australian studies (reported by Rothman, 2002)

Reported SD of 10. Therefore:

$$d = \frac{\overline{X}_F - \overline{X}_M}{10}$$

2. For studies reporting means and standard deviation for males, means and standard deviation for females and number of cases for each sex (e.g. WA monitoring studies, Hogrebe et al., 1985; Johnson, 1973-74)

$$d = \frac{\overline{X}_F - \overline{X}_M}{\sqrt{\left(\frac{(N_F - 1)s_F^{\ 2} + (N_M - 1)s_M^{\ 2}}{N_F + N_M - 2}\right)}}$$

which is the mean for females minus the mean for males divided by the within-group (also called ’pooled’) standard deviation (see Hunter et al., 1982, p. 98).

The reason for using the within-group standard deviation instead of the control-group standard deviation was that the within-group standard deviation had only about half the sampling error of the control-group standard deviation. In addition, Cohen (1988, p. 11) stated that “…the ES index for differences between population means is standardised by division by the common within-population standard deviation.”

The reason for subtracting male mean from female mean was that higher average reading performance was expected for females. As a consequence, positive effect sizes denoted superior performance of females whereas negative effect sizes denoted superior performance of males.

3. For the Botswana study (reported by Fuller et al., 1994)

Using t-test values and number of cases to calculate effect size.

First step: Calculate correlation coefficient r from t-test (see Hunter et al., 1982, p. 98):

$$r = \frac{t}{\sqrt{t^2 + N - 2}}$$

Second step: Calculate effect size *d* based on r:

$$d = \frac{r}{\sqrt{1 - r^2 \times p \times q}}$$

where p is the proportion of females and q is the proportion of males in the sample.

Note that Hunter et al. (1982, p. 98) stated that in the case of equal sample sizes for the two groups "[…] for small correlations, this meant *d*=2r[…]".

4. For studies that record percentages (Gambell and Hunter, 2000; and for NAEP 1971, 1975 and 1980 reported by Levine and Ornstein, 1983)

d=SUM(ASIN(p)-ASIN(q))

5. For the United Kingdom study that reported means for females and males plus the respective standard errors and not the standard deviation

Cohen (1988, p. 6) states

> "..one conventional means for assessing the reliability of a statistic is the standard error (SE) of the statistic. If we consider the arithmetic mean of a variable X ($\overline{X}$), its reliability may be estimated by the standard error (SE) of the mean ($SE_{(\overline{X})}$):"

First, obtain SD from SE:

$$SE_{\overline{X}} = \sqrt{\frac{SD^2}{n}}$$

therefore

$$SE_{\overline{X}} = \frac{SD}{\sqrt{n}}$$

therefore

$$SD = SE_{\overline{X}} \times \sqrt{n}$$

Then, replace SD by $SE_{\overline{X}} \times \sqrt{n}$ into the ordinary formula for *d* to calculate the effect size:

$$d = \frac{\overline{X}_F - \overline{X}_M}{\frac{N_F \times SE_F \times \sqrt{N_F} + N_M \times SE_M \times \sqrt{N_M}}{N_F + N_M - 2}}$$

6. For the NAEP studies mean differences (directly off website)

Reported SD of 50, therefore:

$$d = \frac{\overline{X}_F - \overline{X}_M}{50}$$

7. For PISA 2000 studies

Achievement scores were scaled to a mean of 500 and a standard deviation of 100 (Adams and Wu, 2002). Therefore:

$$d = \frac{\overline{X}_F - \overline{X}_M}{100}$$

8. For studies reporting correlation coefficients (includes the Reading Comprehension Study)

$$d = \frac{r}{\sqrt{1 - r^2 \times p \times q}}$$

where p is the proportion of females and q the proportion of males in the sample.

9. For studies reporting partial correlation coefficients, regression weights, or gammas

These were considered more precise estimates of the relationship between gender and reading achievement as the effects of other variables had been partialled out. In other words, these measures provided information on the strength of the relationship between gender and reading achievement after the influences of other variables on the relationships had been taken into account. In line with this argument, betas or gammas of the most complex models were used as a basis for calculating the effect size as these were considered to be better estimates of the relationships between gender and reading achievement, taking into account the other variables.

In line with Pedhazur (1982) regression coefficients could be considered similar in nature to correlation coefficients. Hence, the same formula as for correlation coefficients was used in the calculation of effect sizes from partial correlations, regression coefficients and gammas (from hierarchical linear models).

10. For studies reporting sum of squares as a result of ANOVA analyses (Oakland and Stern, 1989)

The idea that it was legitimate to use the following formula in calculating effect sizes based on sums of squares was put forward by Keppel (1991, p. 437-444).

$$d = \frac{\sqrt{SS_{Sex}}}{\sqrt{SS_{Total}}}$$

Like partial correlation or regression coefficients, this measure was considered to be better as it took into account other variables, such as Race and SES in the analysis by Oakland and Stern (1989).

11. For the Reading Literacy Study

According to Cohen’s formula (1988, p. 20):

$$d = \frac{\overline{X}_F - \overline{X}_M}{\sigma}$$

whereby values for means for males and females were taken from Purves and Elley (in Elley 1994, p. 106) and the pooled standard deviation for the overall reading score was taken from Elley and Schleicher (in Elley 1994, p. 57).

REFERENCES

* indicates reports from which data were used in the meta analyses.

Adams, R.J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162-172.

Adams, R., and Wu, M. (Eds.). (2002). *Programme for International Student Assessment (PISA): PISA 2000 Technical Report.* Paris: OECD.

Aiken, L.S. and West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA.: Sage.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-1984 technical report.* Princeton, NJ: Educational Testing Service.

Campbell, J.R., Hombo, C.M. and Mazzeo, J. (2000). *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance, NCES 2000–469*, Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics, NAEP. [Online] http://nces.ed.gov/naep3/pdf/main1999/2000469.pdf [Last accessed 24/11/05].

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, (2nd edn.), Hilldale, NJ.: Lawrence Erlbaum Associates.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A. and Mosteller, F. (1992). *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Foundation.

Dedze, I. (1995). *Reading Achievement Within the Educational System of Latvia: Results from the IEA Reading Literacy Study*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 18-22.

*Elley, W.B. (Ed.). (1994). *The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-Two School Systems*. Oxford: Pergamon Press.

*Fuller, B., Hua, H. and Snyder, C.W. Jr. (1994). Focus on gender and academic achievement. When girls learn more than boys: The influence of time in school and pedagogy in Botswana. *Comparative Education Review*, 38(3), 347-376.

*Gambell, T. and Hunter, D. (2000). Surveying gender differences in Canadian school literacy. *Journal of Curriculum Studies*, 32(5), 689-719

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5, 351-379.

Glass, G.V., McGaw, B., and Smith, M.L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, OA.: Sage Publications.

*Glossop, J.A, Appleyard, R., and Roberts, C. (1979). Achievement relative to a measure of general intelligence. *British Journal of Educational Psychology*, 49, 249-257.

*Gorman, T.P., White, J., Orchard, L. and Tate, A. (1982). *Language Performance in Schools. Secondary Survey Report no 1*. Department of Education and Science, Welsh Office, Department of Education for Northern Ireland. London: Her Majesty's Stationery Office.

Gorman, S. (2005). Director for design and analysis, Assessment Division, NCES. Response to the question since when NAEP uses plausible values. Personal e-mail communication via Taslima Rahman 30/11/05.

Hedges, L.V. (1981).Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128.

*Hedges, L.V. and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269, 41-45.

Hedges, L.V. and Olkin, I. (1983). Regression models in research synthesis. *The American Statistician*, 37(2), 137-140.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL.: Academic Press.

*Hogrebe, M.C., Nist, S.L. and Newman, I. (1985). Are there gender differences in reading achievement? An investigation using the high school and beyond data. *Journal of Educational Psychology*, 77(6), 716-24.

Hox, J.J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Hunter, J.E. and Schmidt, F.L. (2004). *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*. (2nd ed.) Thousand Oaks, CA: Sage Publications.

Hunter, J.E., Schmidt, F.L. and Jackson, G.B. (1982). *Meta-Analysis. Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage.

Husén, T. (1967). *International Study of Achievement in Mathematics. A Comparison of Twelve Countries. Volume II*. Stockholm: Almqvist and Wiksell.

*Johnson, D.D. (1973-1974). Sex differences in reading across cultures. *Reading Research Quarterly*, 9(1), 67-86.

Johnston, J. and Dunne, M. (1996). Revealing assumptions: Problematising research on gender and mathematics and science education. In L.H. Parker, L.J. Rennie, B.J. Fraser (Eds.) *Gender, science and mathematics. Shortening the shadow* (p. 53-63). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice Hall.

Keeves, J. P. (1988). Sex differences in ability and achievement. In J.P. Keeves (Ed.) *Educational research, methodology, and measurement: An international handbook* (pp. 689-700). Oxford: Pergamon Press.

Kulik, J.A., and Kulik, C.-L.C. (1989). Meta-analysis in education. *International Journal of Educational Research*, 13(3), 221-340.

*Levine, D.U. and Ornstein, A.C. (1983). Sex differences in ability and achievement. *Journal of Research and Development in Education*. 16 (2), 66-72.

Lietz, P. (1996). *Reading Comprehension Across Cultures and Over Time*. Münster/New York: Waxmann.

Lietz, P. (in press). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation*.

Light, R.J. and Smith, P.V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41(4), 429-471.

Marks, G.N. and Ainley, J. (1996). *Reading Comprehension and Numeracy Among Junior Secondary School Students in Australia. Longitudinal Surveys of Australian Youth, Research Report Number 3*. Melbourne: Australian Council for Educational Research.

*Neuman, S.B. and Prowda, P. (1982). Television viewing and reading achievement. *Journal of Reading*, 25, 666-670.

*Oakland, T., and Stern, W. (1989). Variables associated with reading and math achievement among a heterogeneous group of students. *The Journal of School Psychology*, 27, 127-140

Pedhazur, E.J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*. (2nd ed). Fort Worth, TX: Holt, Rinehart and Winston.

*Plisko, V.W. (2003). *The Release of the National Assessment of Educational Progress (NAEP) The Nation's Report Card: Reading and Mathematics* 2003. [Online] http://nces.ed.gov/commissioner/remarks2003/11_13_2003.asp [Last accessed 18/05/05].

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S.W. and Bryk, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10(2), 75-98.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F. and Congdon, R. (2001). *HLM 5. Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: SSI Scientific Software International.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage Publications.

*Rothman, S. (2002). *Longitudinal Surveys of Australian Youth. Research Report Number 29. Achievement in Literacy and Numeracy by Australian 14-Year-Olds, 1975-1998*. Hawthorne, Vic.: Australian Council for Educational Research. [Online] http://www.acer.edu.au/research/projects/lsay/reports/lsay29.pdf [Last accessed 24/11/05].

*Shilling, F., and Lynch, P.D. (1985). Father versus mother custody and academic achievement of eighth grade children. *Journal of Research and Development in Education*. 18(2), 7-11.

Slavin, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher*, 13, 6-15.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher*, 15, 5-11.

Thorndike, E.L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. Reprinted in *Reading Research Quarterly*, (1971), 6(4), 425-434.

Thorndike, R.L. (1973). *Reading Comprehension Education in Fifteen Countries. International Studies in Evaluation III*. Stockholm: Almqvist and Wiksell.

Tracy, D.M. (1987). Toys, spatial ability, and science and mathematics achievement – Are they related? *Sex Roles*, 17(3-4), 115-138.

*U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP), 2003, 2002, 1998, 1994 and 1992 Reading Assessments. NAEP Data Tool v3.0*. [Online] http://nces.ed.gov/ nationsreportcard/naepdata/getdata.asp, search options “subject”= reading; “Grade”=Grade 8 and Grade 12; “State/Jurisdiction”=Nation; Category=Major reporting groups, after pressing “continue” select “gender”. [Last accessed 06/06/05].

Wagemaker, H., Taube, K., Munck, I., Kontogiannopoulou-Polydorides, G. and Martin, M. (1996). *Are Girls Better Readers? Gender Differences in Reading Literacy in 32 Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Warm, T.A. (1985). *Weighted Maximum Likelihood Estimation of Ability in Item Response Theory with Tests of Finite Length* (Technical Report CGI-TR-86-08). Oklahoma City: U.S. Coast Guard Institute.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128

*Youngman, M.B. (1980). Some determinant of early secondary school performance. *British Journal of Educational Psychology*, 50, 43-52.

❖IEJ