

Chapter 21

THE USE OF META-ANALYSIS IN CROSS-NATIONAL STUDIES

EDITH D. DE LEEUW
JOOP J. HOX

21.1 INTRODUCTION

Meta-analysis offers excellent tools for researchers in the field of cross-national and cross-cultural studies. It provides researchers with methods to combine the outcomes of different studies and analyze the results to investigate potential differences between countries or cultures. Meta-analysis can help those conducting cross-national and cross-cultural research and users of international databases decide if results are comparable across countries and thus unravel general and country-specific information.

Meta-analysis has its origin and some of its earliest applications in the educational sciences. The term 'meta-analysis' was introduced by Glass (1976) in his presidential address to the American Educational Research Association (AERA) as "... the analysis of the results of statistical analysis for the purpose of drawing general conclusions." One of the first meta-analyses published was on the effect of class size on educational achievement (Glass and Smith 1979). Meta-analysis was adopted quickly in psychology and bio-medical research, and soon meta-analyses were published on topics as varied as the effectiveness of psychotherapy (Smith, Glass, and Miller 1980) and the effect of aspirin on the occurrence of heart attacks (Barnett et al. 1988). Interest in meta-analysis increased dramatically in the eighties and nineties. The topics broadened from summary studies estimating an overall effect (e.g., does class size affect achievement) to studies focused on differential effects for different cultural or ethnic subgroups. Examples are studies on cultural differences in child competitiveness (Strube 1981), gender and cognitive performance (Signorella and Jamieson 1986), and race effects in performance evaluation (Ford, Kraiger, and Schechtman 1986). Interest in using meta-analysis for theoretical explanation rather than summary description is now increasing (Cook et al. 1994).

Despite its name, meta-analysis is not a single type of method or analysis but a set of methods: a methodology for the systematic combination of information from several different sources. In other words, meta-analysis is a systematic approach for integrating the outcomes of a set of studies, summarizing what is common and analyzing what is different. It started as a formal method for systematic literature review. The medical sciences used meta-analysis to combine the results of so-called

'multi-site' experiments or clinical trials, thereby establishing an important second use of the techniques. At present, there are two important applications of meta-analysis: (1) systematic literature review, and (2) systematic analysis of data collected at different sites. Both approaches are valuable in cross-national and cross-cultural studies. Below, we describe these two approaches then discuss the application of meta-analysis in cross-cultural and cross-national research.

21.2 TWO APPROACHES TO META-ANALYSIS

21.2.1 Meta-analysis for Literature Review Summaries

Meta-analysis started as a method for reviewing research literature. The traditional narrative literature reviews no longer met the needs of researchers (Glass, Smith, and McGaw, 1981; Rosenthal 1984). Narrative reviews were considered subjective and an inefficient way to extract useful information from the literature (Light and Pillemer 1984). Quantitative procedures therefore were developed to integrate the outcomes of individual studies. This approach prescribes formal methods not only for the statistical summary of results, but also for the earlier stages, such as collecting the relevant literature and coding the results of the individual studies. These formal methods need to be described clearly in the meta-analysis report, in order to allow the analysis to be replicated and evaluated (Light and Pillemer 1984; Cooper and Hedges 1994).

Problem Formulation

The first step in a scientific meta-analysis is the *problem formulation*. One should start with a clear description of the research problem, including the universe to which one wants to generalize. For example, in a medical application, such as the Barnett et al. (1988) aspirin study, the typical problem formulation would focus on the effectiveness of the intervention: does aspirin help to prevent heart attacks? In other applications, the focus may be on the role of moderator variables. For example, when a researcher is interested in the role of incentives in raising survey response rates, she wants not only to know if incentives have worked in the past (descriptive meta-analysis) but also to be able to generalize to future and probably somewhat different studies. Thus, while the basic question is still whether incentives will increase response rates in later surveys, additional research questions might include the following: "Does the effect vary by mode of data collection?", "What is the best strategy: prepaid or promised incentive?", "What type of incentive works better: money or gift?" Here the interest is not only in the combined outcome but also in study-level explanatory variables that moderate the effect of incentives on response rates (see Singer et al. 1999 for a good example).

In comparative meta-analyses, the most important aspects of the problem formulation are explicit research questions on differences between countries and/or cultures within countries; examples are Daly (1996), Schimmack (1996), and Van de

Vijver (1997). Daly's research question is whether companies act in manners consistent with behaviorist or cognitive organizational theories. She compares American and international studies to find out if there are intercultural differences. Her findings are that companies tend to behave in accordance with cognitive theories, with no discernable intercultural differences. Schimmack investigates whether the structure of the correlations between the recognition of facial expressions of emotions is equivalent across different countries. He concludes that there may be real cross-cultural differences with respect to the recognition of sadness and fear. Finally, Van de Vijver presents a comparative meta-analysis of cross-cultural comparisons of cognitive test scores. Again, the focus of the meta-analysis is on explaining cross-cultural differences. His comparative meta-analysis illustrates that a meta-analysis does not need to be restricted to simple research questions. He investigates several models that may explain cross-cultural differences in cognitive test performance. He finds that differences in cognitive performance are positively related to the degree of affluence of cultural groups and that these differences increase with age and education. The performance differences are larger on common Western tasks and smaller on locally developed non-Western tasks. There were no differences in abstract thinking. Finally, only intranational studies show a relation with the complexity of the task. In general, he finds that intranational comparisons are related to different explanatory variables than are cross-national comparisons.

Retrieval of Relevant Literature

The second step is the data collection phase: *retrieval of literature*. In this phase, the researcher conducts a literature search for relevant studies. Based on the research question of interest, the researcher has to formulate a search strategy and define key concepts and key words. This problem is analogous to the problem of sampling respondents in a survey. The researcher must begin by defining the universe of interest. In meta-analysis, this involves decisions about the time period to be searched (e.g., are older publications still relevant?), whether *all* studies are to be included or only studies that meet certain methodological requirements (e.g., do we include nonrandomized experiments or trials?), and whether unpublished and 'gray' literature is included. After defining the population of studies, a search strategy is defined, based on the research question of interest, and key concepts and key words are defined. Reference databases are a good starting point. There are specialized databases on CD-ROM for each field of research (e.g., AgeLine for gerontology, DRUGINFO for interdisciplinary research on substance abuse, ERIC for the educational sciences, MEDLINE for healthcare, PsycINFO for psychology, Sociological Abstracts for sociology, SRM for methodology in the broadest sense).

A generic problem with computerized databases is that they are compiled by information scientists who cannot know the problem formulation of the actual meta-analysis, so the databases are not tailored to the needs of a given meta-analysis. Typically, these databases focus almost exclusively on literature published in English, and mostly in American research journals. Special efforts therefore are

needed to avoid the resulting Anglo-Saxon selection bias, perhaps especially for comparative research. Thus, it is wise to search several sources. Studying the indices of specialized journals and the references of earlier review papers provides an additional check on the effectiveness of the search. The increased availability of Internet-based resources makes Internet searches a valuable resource, especially for non-U.S. based studies. Good additional strategies are appeals in newsletters and on pertinent e-mail discussion lists, formal appeals to scholars in the field, and in informal conversations at conferences. Since not all countries share the strong Anglo-Saxon pressure to 'publish-or-perish' in international journals, it is important to search for unpublished reports, conference papers, and recent studies (the gray literature). Fink (1998) provides suggestions for such 'multiple approach' strategies.

Coding of the Studies

Systematic coding of the studies: The goal of this step is to transform the corpus of studies into a data matrix. Each row then represents a study or a case in statistical terms and each column a variable, such as study characteristics and outcome variables. A detailed coding schedule is needed, which also must be described in the methods section of the meta-analytic report; this is often added as an Appendix.

A good starting point for designing a coding scheme is to inspect schemes used in other meta-analyses on related topics. For example, de Leeuw (1992) took a coding scheme from Sudman and Bradburn (1974) as the inspiration and starting point for her meta-analysis of the influence of data collection method on data quality. The general study characteristics, background variables, and the outcome variables of interest are coded. Usually, part or all of the coding is carried out by two independent coders, and a measure of intercoder reliability (e.g., Cohen's kappa) is reported.

Study characteristics and background variables. In meta-analysis, background variables that relate to the research report are coded, such as year of publication, country where the study was done, and journal or type of publication (e.g., dissertation, conference paper). In addition, meta-analysis codes characteristics that describe the study and its design. For instance, in de Leeuw's (1992) meta-analysis on data collection modes, the study characteristics included type of sample, sample size, topic of survey, saliency, and equivalence of questionnaire in different modes.

A special background or study characteristic is 'quality of study'; this is often a general evaluation based on coded study characteristics, such as experimental design (e.g., true random experiment or not, control for attrition, use of placebo, a double blind treatment in medical research, etc.; cf. Wortman 1994). A good example is the meta-analysis by Wortman and Bryant (1985). They criticized earlier meta-analyses on the effect of desegregation on educational attainment of minority groups in the United States that suggested that desegregation had beneficial effects. These early meta-analyses included all the studies found, including studies with relatively weak research designs. In response to the ensuing debate, Wortman and Bryant reanalyzed the studies *but* included methodological criteria in the analysis. Studies that fell

below a predefined quality level were excluded. Wortman and Bryant concluded that methodologically strong studies tended to find a smaller effect of desegregation. This smaller effect in the stronger studies was nonetheless large enough to be of substantive importance and to have valid implications for educational policy.

Outcome variables. Attention in meta-analysis focused first on statistical significance and the combination of significance levels (*p*-values). The main outcome variable coded in this case was the *p*-value of the statistical test in the original publications. The statistical combination of *p*-values (Becker 1994) focuses on the research question "Is there a statistically significant effect?" Soon attention in meta-analysis shifted to the effect size (see Cohen 1969, 1988) and methods for combining effect sizes. This approach focuses on the research question "How large is the effect?"

There are several formulae for effect size, and the specific effect size chosen depends on the research question and the design of the original studies. For example, for experimental studies the standardized mean difference between the experimental and control group (the difference between the group means divided by their common standard deviation) is a common estimator of effect size:

$$d = (\bar{x}_E - \bar{x}_C) / s. \quad (21.1)$$

In correlational research, the usual effect size is the correlation coefficient. For studies that rely on categorical outcome variables, effect size estimates use proportions—for example, the relative risk:

$$RR = P_E / P_C, \quad (21.2)$$

is defined as the ratio of the proportion of a specified outcome in the experimental and control group (cf. Cornell and Mulrow 1999).

The use of effect size estimators in meta-analysis is based on the assumption that the studies all estimate the same parameter and that this parameter of interest can be defined using a common metric (e.g., standardized mean difference, correlation, relative risk). The goal of the meta-analysis is then to combine all effect sizes into one overall 'superoutcome' and at the same time give a description of its sampling variability. In the end, the meta-analyst reports an estimate of the overall effect size and a confidence interval, considering all available information.

A good example of combining outcomes is de Leeuw's (1992) meta-analysis of the differences between mail, telephone, and face-to-face surveys. She found 67 articles and papers that compared at least two of these data collection modes on data quality. Three articles reanalyzed earlier studies, one study contained severe design flaws, and ten articles did not provide enough information for coding. In the end, 52 studies were available. The quality criteria coded were response validity, item nonresponse, number of statements in reply to open questions, social desirability, and similarity of response distributions. Not all the studies compared all three modes, and not all

presented results for all quality criteria. This is typical for a meta-analysis involving several comparisons and multiple outcome variables. De Leeuw (1992) presents the results in the form of pairwise comparisons of the modes, using correlation coefficients as the common effect size measure. For example, for social desirability, comparing face-to-face and telephone surveys, she finds a mean $r = -0.01$ (95% confidence interval: -0.03 to $+0.01$), based on 14 studies, and comparing mail and face-to-face surveys, she finds a mean $r = +0.09$ (95% confidence interval: $+0.07$ to $+0.11$), based on 13 studies. Finally, comparing mail and telephone surveys, she finds a mean $r = +0.06$ (95% confidence interval: $+0.03$ to $+0.09$), based on five studies. De Leeuw concludes that concerning social desirability bias, mail surveys perform somewhat better than face-to-face and telephone surveys, which do not differ from one another in this respect.

Statistical Analysis of the Coded Data

The fourth step in meta-analysis is *statistical analysis of the coded data*. Basic statistical techniques are used to summarize the results. More sophisticated techniques are used to investigate potential heterogeneity in the data. The latter are very important in cross-national studies, as they address the question of whether the outcomes of the studies are comparable across countries.

The basic statistical approaches are (1) summarize p -values (summarize significance levels of individual studies) and (2) summarize outcomes. There are many methods to *combine p -values* (Becker 1994). One of the most robust is Stouffer's method. In this, each p -value is first transformed to a z -score via a standard normal transformation, then these z -scores are summed up across all studies and the sum is divided by the square root of the number of p -values. Finally, this overall z -score is transformed back to a p -value. If the combined p -value is smaller than a chosen significance level, the null hypothesis of no effect is rejected, and one may conclude that there is an effect in the studies investigated.

For each effect size estimator, there is a statistical procedure to *combine the effect sizes* into one overall effect size and calculate the corresponding standard error (Hedges and Olkin 1985). We illustrate these procedures to combine effect size with an example of the combination of an effect size d (e.g., the standardized difference between females and males in reasoning) over studies. The statistical procedure has two important assumptions. First, one assumes that all studies estimate the same effect (e.g., all studies study a homogeneous domain such as reasoning, and *no other* type of intelligence test). Second, one assumes that the outcomes are homogeneous, that is, that the variation in study outcomes is exclusively attributed to sampling variance and that there is *no systematic variance* associated with other sources. Each study coded provides an estimate of the effect size d_i and an accompanying standard error se_i . These are statistically combined into one overall outcome $d = \delta$ with corresponding standard error, which are used to decide whether there is a statistically significant effect over studies and to estimate a confidence interval.

To combine the effect sizes, a weighted integration method is used, which weights each study with the inverse of its sampling variance given by:

$$w_i = 1/se_i^2. \quad (21.3)$$

Thus, the weighted integration method estimates the combined effect size as

$$\bar{d} = \sum w_i d_i / \sum w_i. \quad (21.4)$$

The basic meta-analytical statistics above follow the *fixed effect model*, which is only valid under the assumption of *homogeneity*. Homogeneity means that all variance between studies is only sampling variance. If there are real differences between the studies, other sources contribute to the variance, and the fixed effect model is no longer valid. This means that the *random effects* model should be adopted, which includes this extra source of variation in the appropriate statistical formulas. This is obviously of central relevance for cross-cultural studies, and we discuss it in more detail in section 21.3. First, we describe another very useful application of meta-analysis: the combination of data from multi-site studies.

21.2.2 Meta-analysis for the Combination of Data: Multi-Site Studies

Especially in the biomedical sciences, meta-analysis is used not only for combining existing results but also in the design of prospective studies. In medical research, the number of cases available at each site (hospital) is often too small to permit the essential analyses. Therefore, data from different sites are combined using meta-analytic procedures. Essential in such *multi-site studies* is that all the experiments are replications, investigating the same research question and using the same methods. To achieve comparability, an explicit protocol is used that describes all the important aspects of the study's design and data collection procedures. The data are collected in medical practices and hospitals located in different places all over the world. Since populations may be different at different sites and organizational and cultural differences between the sites may necessitate small deviations in the protocol, background variables that describe the site and the procedures involved may be used to resolve uncertainties when such deviations occur.

Richard Peto pioneered prospective uses of meta-analysis for clinical trials. He persuaded 5,000 British doctors to participate in a multi-site clinical trial on the effectiveness of aspirin in preventing heart attacks. The Cochran Collaboration is another well-known international project, supported by 15 medical centers worldwide (Cornell and Mulrow 1999, see also <http://hiru.mcmaster.ca/cochrane>).

In international studies, it is essential that collateral background information on countries involved is collected and coded. For example, different countries may differ in health regulations or dietary customs, and protocols on how to treat

'dropout' may differ between sites. In statistical terminology, there are potential sources for heterogeneity, a situation also common in cross-cultural and cross-national studies. In meta-analysis, this is solved by coding available background variables and modeling the heterogeneity. In prospective meta-analysis, the need for background variables to explain differences between countries should be anticipated by including the collection of important background data in the study design.

21.3 STATISTICAL METHODS TO ANALYZE HETEROGENEITY

The goal in meta-analysis is to summarize the results of many studies, preferably in one clear 'summary' outcome. To justify this, *homogeneity* must be assumed: all studies must estimate the same fixed parameter, and all variance is assumed to be only sampling variance. If the study outcomes are heterogeneous, a simple fixed effect model to summarize the outcomes is no longer valid. For instance, if study outcomes differ for different countries, the country of origin is an additional source of variation. This additional, systematic, variation should be included in the statistical model; a *random effects* model should be adopted. In view of their relevance for comparative research, heterogeneity and the random effects model are discussed in more detail.

We illustrate the statistical procedures for meta-analysis and tests for homogeneity by analyzing a small and manageable data set that illustrates the main points to be made about meta-analysis in general. We follow classical meta-analysis methods as implemented in the program META (Schwarzer 1989). The data set consists of six studies assessing the effect of taking aspirin after a heart attack comparing an experimental group and a control group which was administered a placebo (Draper et al. 1992). The data are presented in Table 21.1. Note that study six has a strongly significant effect in the opposite direction, which results in a very high one-sided p -value. This could be an indication of heterogeneity, and that can be tested using a formal chi-square test for homogeneity.

Table 21.1. Mortality Rates for Aspirin and Placebo Control Group

| Study | Aspirin (E) | | Placebo (C) | | Comparison | | |
|-------|-------------|-----------|-------------|-----------|------------|-------------|-------|
| | N_E | Mortality | N_C | Mortality | d | s.e.(d) | p |
| 1 | 615 | .0797 | 624 | .1074 | -.1666 | .0569 | .0017 |
| 2 | 758 | .0580 | 771 | .0830 | -.1866 | .0513 | .0001 |
| 3 | 317 | .0852 | 309 | .1036 | -.1096 | .0800 | .0853 |
| 4 | 832 | .1226 | 850 | .1482 | -.1179 | .0488 | .0079 |
| 5 | 810 | .1049 | 406 | .1281 | -.1187 | .0609 | .0256 |
| 6 | 2267 | .1085 | 2257 | .0970 | +.0643 | .0297 | .9847 |

We use the standardized effect size d , which for the comparison of two proportions is calculated as

$$d = Z_{p_E} - Z_{p_C} \quad (21.5)$$

where Z_p is the inverse of the standard normal distribution corresponding to the proportion p (Hedges and Olkin 1985). The sampling variance of d is $(n_E + n_C) / (n_E n_C) + d^2 / 2(n_E + n_C)$, where n_E and n_C are the sample sizes of the experimental and control groups. The p -values in Table 21.1 are the left-sided p -values for $Z = d / \text{s.e.}(d)$.

21.3.1 Classical Meta-analysis

Classical meta-analysis contains a variety of complementary approaches. One simple approach combines the p -values of the studies into one overall p -value for the collection of studies. For the aspirin example, the Stouffer method gives a combined Z of 4.17, with $p < .001$.

The combined p -value gives us proof that an effect exists but no information on the size of the experimental effect. The next step combines the effect sizes of the six studies into one overall effect size and establishes the significance or a confidence interval for this combined effect. Using a fixed effect model, the weighted integration method estimates the combined effect size as -0.06 , with a 95% confidence interval extending from -0.10 to -0.02 . However, the deviant outcome of Study 6 strongly suggests that the effects are heterogeneous, that is, that the effects differ across studies. In this event, the random effects model is more appropriate for combining the studies. The usual homogeneity test is a chi-square test on the residuals, which for our example leads to $\chi^2 = 30.2$, $df = 5$, $p < .001$. This is significant, so we conclude that the outcomes are strongly heterogeneous.

We used a formal homogeneity test to decide whether a fixed or a random effects model is appropriate. Hunter and Schmidt (1990) propose an additional criterion: the size of the between-studies variance. They suggest that if this is more than 25% of the total variance, it is important. In our example, the proportion of systematic between-studies variance v_θ is estimated as 0.60, much larger therefore than the lower limit of 0.25 that Hunter and Schmidt propose.

Since the between-studies variance is large and significant, random effects meta-analysis must be used. First, the between-studies variance v_θ is estimated. This is added to all the individual variances, so the estimated variance for each study becomes $v_i^* = v_\theta + v_i$. This leads to different weights and different estimates.

Classical meta-analysis uses a simple estimate for the between-studies variance (Lipsey and Wilson 2001, 119). This leads to an overall effect estimate of $\delta = -0.10$, with a 95% confidence interval from -0.17 to -0.03 . These results differ from the results of the fixed effect model, because the random effects model takes into account the systematic variation between the studies when the studies' outcomes are

averaged. The interpretation of the random effects estimates is also different. The overall effect size of -0.10 is the average of the distribution of effects in the population of studies. The average outcome is negative, that is, the aspirin group has significantly fewer heart attacks than the placebo group.

When the outcomes are heterogeneous, the classical approach is to divide the studies into clusters that have different average effect sizes but are internally homogeneous. In our example, this cluster analysis produces two clusters. The first consists of Studies 1 through 5, and the second consists of Study 6. In the first cluster, the variances are homogeneous, which means across these five studies there is no systematic between-study variance. The lack of available background information prevents further analysis of the aspirin data, a problem typical for meta-analysis (Lipsey and Wilson 2001).

In many cases, it would be impractical for primary researchers to publish their raw data. There is a real need for a publication practice that encourages researchers to publish sufficient statistics and their corresponding standard errors for various subgroups. The American Psychological Association already supports such publication practices and recommends that researchers retain their raw data for possible use by others (APA 1994). Developments in data archiving and using the Internet also will help.

21.3.2 Multilevel Meta-analysis

Multilevel regression analysis can be used to estimate the random effects model for meta-analysis, including available explanatory variables (Raudenbush 1994; Hox 2002). In the multilevel approach to meta-analysis, we recognize the two-level structure that is implicit in Table 21.1. We have six studies, with a total of 10,816 patients. If we had access to the raw data of all studies, we could set up a standard multilevel regression analysis with patients at the lowest level, studies at the highest level, and explanatory variables that describe patient or study characteristics. Since we do not have access to the raw data, a special multilevel model is used instead (Kalaian and Raudenbush 1996; Hox 2002).

Multilevel regression analysis is a regression model with a complicated error structure. For an introduction to multilevel modeling, see Hox (1995, 2002). Since we have no access to the raw data, we set up a model for the sufficient statistics: the d 's and their standard errors. In such a model, the lowest level sampling variance is not estimated, because it is known from the sampling errors provided in Table 21.1. The term 'variance known' model (Bryk and Raudenbush 1992) aptly describes this characteristic feature of multilevel meta-analysis.

A multilevel meta-analysis of the six aspirin studies produces virtually the same results as the classical meta-analysis reported above. The starting model is a model without explanatory variables. The intercept, which is the overall outcome, is estimated as -0.10 , with a 95% confidence interval from -0.18 to -0.02 . The null hypothesis of homogeneous outcomes is rejected. The chi-square test on the

significance of the between-studies variance yields a chi-square of 34.34 ($df = 5$, $p < .001$). The proportion of systematic variance is estimated as 0.73, a bit higher than in the classical meta-analysis.

The power of multilevel meta-analysis becomes clear when we model the differences in the study outcomes. We have no real background variables that describe the studies. Given the deviant outcome of Study 6, we decide to model the heterogeneity by including a dummy variable that represents Study 6 as an explanatory variable in the model. The regression coefficient for this variable is 0.21 ($p < .01$), which means that in Study 6, the difference between the experimental and control group is 0.21 larger than in the other studies but in an unexpected direction. The intercept, which in this model is the average outcome of the other five studies, is -0.14 , with a standard error of 0.025 ($p < .001$). After including the dummy variable as an explanatory variable in the model, the residual between-studies variance is no longer significant ($\chi^2 = 1.50$, $df = 4$, $p = .83$). Thus, the heterogeneity is completely explained by the deviant outcome of Study 6.

In contrast to the cluster analysis reported above, we now have a formal hypothesis test for the hypothesis that Study 6 is an outlier and for the assertion that after controlling for this variable, there is no between-study variance left. If we had had more background variables that code for study characteristics, we could have added them to the model in the same manner.

Table 21.2 presents the results from the analyses of the aspirin data. The fixed effect model yields the smallest estimate of the combined effect \bar{d} . The classical random effects model and the multilevel meta-analysis null-model are equivalent, but the estimation procedures differ, since multilevel analysis uses Maximum Likelihood estimates. The differences are small but noticeable. In the multilevel model that includes a dummy for the unusual Study 6, the variance between the studies is estimated as zero. In this model, the between-studies variance is an estimate of the residual variance, and the chi-square test is a test for the significance of this residual variance. Thus, after taking into account the distinction between Study 6 and the others, there is no between-study variance left.

Table 21.2. Results from Classical and Multilevel Meta-analysis on Aspirin Data

| Model: | Fixed Effect | | | Random Effects | | | Multilevel Null Model | | | Multilevel with Study 6 Dummy | | |
|-----------------------|--------------|--------|-----|----------------|--------|-----|-----------------------|--------|-----|-------------------------------|--------|-----|
| | \bar{d} | (s.e.) | p | \bar{d} | (s.e.) | p | coef. | (s.e.) | p | coef. | (s.e.) | p |
| Intercept | -.06 | (.019) | .00 | -.10 | (.035) | .00 | -.10 | (.043) | .02 | -.14 | (.035) | .00 |
| Study 6 | N/A | | | N/A | | | - | | | +21 | (.039) | .00 |
| Study var. | N/A | | | .005 | | | .008 | | | .000 | | |
| χ^2 (df) p | 30.2 | (5) | .00 | 30.2 | (5) | .00 | 34.3 | (5) | .00 | 1.5 | (4) | .83 |

21.4 META-ANALYSIS IN COMPARATIVE RESEARCH

The following section describes two different meta-analyses of cross-national data sets that illustrate important issues in cross-national comparisons. The first example is a cross-national literature review and the second a cross-national multi-site study.

21.4.1 Multicultural and Multinational Approaches in Literature Reviews

The goal of meta-analysis for literature reviews is generalization over studies to summarize outcomes. In cross-cultural and cross-national research, we routinely expect heterogeneous results. The modern approach to heterogeneous results is to investigate if explanatory variables influence this general outcome. This is especially important in cross-cultural and cross-national comparisons. In these comparisons, it is important to control for any differences in the data collection methods and instrumentation used in the different countries. In addition, we need explanatory variables to explain real conceptual differences between different countries. Thus, testing and modeling heterogeneity become the central issue. This makes multilevel meta-analysis a powerful tool in cross-cultural and cross-national meta-analyses, because testing and modeling heterogeneity is an integral part of multilevel modeling.

A study by de Leeuw and de Heer (2001) on international nonresponse trends illustrates this. The study combines the results from reports on nonresponse in 16 different countries. Summary data were solicited from the official statistical offices on the response rates of a number of surveys over as many years as were available. Information about the sampling design, survey design, fieldwork strategy, interviewer corps, and survey climate also were collected. Three outcome variables—proportion response, proportion refusals, and proportion noncontacts—were used to investigate the nonresponse process in some detail. The study addresses three research questions: (1) Does nonresponse differ between countries? (2) Does nonresponse increase over time? (3) Is the increase different between countries?

De Leeuw and de Heer did not have raw data at their disposal. Thus multilevel meta-analysis was used to analyze similarities and differences across the 16 countries. The authors find that countries differ significantly in response rate, noncontact rate, and refusal rate, and that type of survey influences response rate and refusal rate but does not influence noncontact rate. There were strong trends over time, but these were not the same for all countries, as Figure 21.1 from de Leeuw and de Heer (2001) clearly shows. De Leeuw and de Heer conclude that countries differ in response rate and that response rates have indeed been declining over the years. However, the trends differ across countries, and the differences in response trends are caused by differences between countries in the rate at which refusals are increasing. Some design and fieldwork factors appear to have an effect.

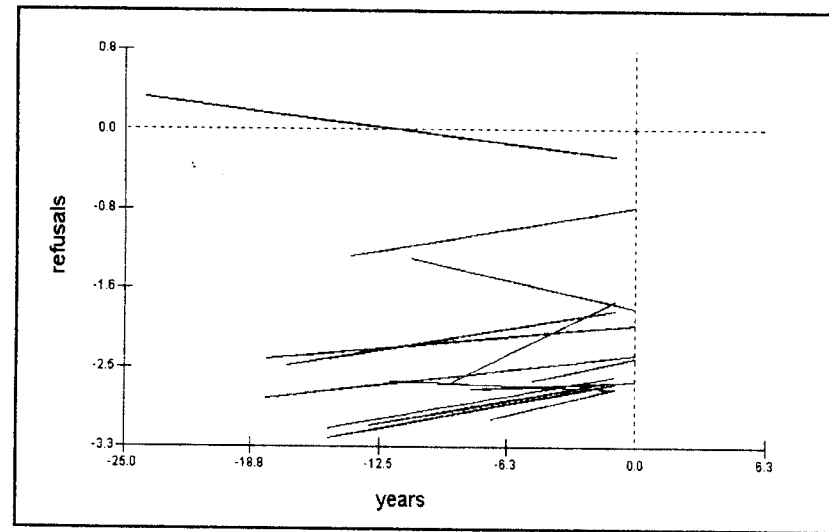


Figure 21.1 Refusals Across Years for Different Countries (1998 = 0; logit scale).

De Leeuw and de Heer's study illustrates some problems typical for meta-analysis. Ideally, to compare nonresponse trends internationally, the data set should contain long and detailed time series, covering a range of survey types and organizations, with all countries providing data for all surveys. In fact, in their data set, countries differ in the surveys they report on, in the interval between two subsequent surveys, and in the length of the time series. To cope with these problems, de Leeuw and de Heer carry out several analyses, using different subsets of the data. They use multilevel regression analysis, which allows inclusion of multiple control and explanatory variables. By comparing different models, they show that their findings point to real differences between the countries.

21.4.2 Meta-analysis to Combine Multi-Site Results from Cross-national Studies

In biomedical research, multi-site studies are an accepted tool to collect data in different places to achieve a sufficient sample size. The goal of such studies is to combine data into one outcome, which assumes homogeneity. In cross-national studies, the interest is often different; heterogeneity is routinely expected, and if found, one would want to explain the heterogeneity. Good examples are the large studies initiated by the International Association for the Evaluation of Educational Achievement (IEA) to compare educational achievement across countries. Here, the interest is squarely on differences between countries and their potential causes.

Two types of variables are important in a comparative study: methodological variables and conceptual or theoretical variables. Methodological variables are used to control statistically for differences in methods between the countries, such as differences in response or data collection methods. The primary aim of introducing methodological variables is to filter out 'method variance' or systematic error. The conceptual variables are used to investigate real differences between the countries, such as differences in educational systems or hours devoted to teaching.

Table 21.3. Reading Achievement Data from 27 Countries

| Country | Reading | s.e. | Age | Grade | Economic | Health | Literacy |
|-----------------|---------|------|------|-------|----------|--------|----------|
| Finland | 569 | 3.40 | 9.7 | 3 | .76 | .52 | 1.26 |
| U.S.A. | 547 | 2.80 | 10.0 | 4 | 1.25 | .03 | .31 |
| Sweden | 539 | 2.80 | 9.8 | 3 | 1.56 | .83 | 1.17 |
| France | 531 | 4.00 | 10.1 | 4 | .55 | .46 | .09 |
| Italy | 529 | 4.30 | 9.9 | 4 | .02 | .18 | -.37 |
| N. Zealand | 528 | 3.30 | 10.0 | 5 | -.43 | .31 | .53 |
| Norway | 524 | 2.60 | 9.8 | 3 | 1.34 | .83 | 1.26 |
| Iceland | 518 | 3.26 | 9.8 | 3 | .69 | 1.04 | 1.29 |
| Hong Kong | 517 | 3.90 | 10.0 | 4 | -.62 | .75 | -.74 |
| Singapore | 515 | 1.00 | 9.3 | 3 | -.50 | -.28 | -.76 |
| Switzerland | 511 | 2.70 | 9.7 | 3 | 2.15 | .61 | 1.10 |
| Ireland | 509 | 3.60 | 9.3 | 4 | -.57 | .37 | .03 |
| Belgium/Fr | 507 | 3.20 | 9.8 | 4 | .39 | .31 | .18 |
| Greece | 504 | 3.70 | 9.3 | 4 | -1.07 | .40 | -.74 |
| Spain | 504 | 2.50 | 10.0 | 4 | -.80 | .75 | -.65 |
| Germany/W | 503 | 3.00 | 9.4 | 3 | .76 | .31 | .59 |
| Canada/BC | 500 | 3.00 | 8.9 | 3 | 1.00 | .40 | .29 |
| Germany/E | 499 | 4.30 | 9.5 | 3 | -.19 | -.21 | 1.37 |
| Hungary | 499 | 3.10 | 9.3 | 3 | -1.14 | -1.54 | .26 |
| Slovenia | 498 | 2.60 | 9.7 | 3 | -.97 | -.43 | -.84 |
| The Netherlands | 485 | 3.60 | 9.2 | 3 | .43 | .83 | .48 |
| Cyprus | 481 | 2.30 | 9.8 | 4 | -.90 | .14 | -.95 |
| Portugal | 478 | 3.60 | 10.4 | 4 | -1.16 | -.49 | -1.65 |
| Denmark | 475 | 3.50 | 9.8 | 3 | .88 | .09 | .63 |
| Trinidad | 451 | 3.40 | 9.6 | 4 | -.81 | -1.09 | -.35 |
| Indonesia | 394 | 3.00 | 10.8 | 4 | -1.52 | -3.77 | -2.70 |
| Venezuela | 383 | 3.40 | 10.1 | 4 | -1.09 | -1.32 | -1.08 |

Table 21.3 presents the results of the IEA study into reading achievement in 27 countries, assembled from information given by Elley (1992, Table 2.1). We have the mean reading achievement of nine-year-old school children, the associated standard error (adjusted for complex survey design), the mean age when tested, the grade when tested, and three indicators of the countries' economic, health, and literacy status. For convenience, the data are sorted by average reading score of each country.

Although the IEA went to great effort to insure that data collection methods were comparable across countries (Elley 1992), Table 21.3 makes clear that children were tested at slightly different ages and in different school grades. These two variables are clear examples of methodological variables. Thus, before comparing countries on reading achievement, we first must adjust the reading scores for differences in mean age and school grade.

Table 21.4 shows the results of three multilevel meta-analysis models on these data. For each model, the table presents the regression coefficient(s) of the explanatory variables in the model, the corresponding standard error (in parentheses), and the resulting *p*-value. In addition, for each model, the between-country variance is given, as is the value of the chi-square for its significance test.

The first model in Table 21.4 is a model without explanatory variables. Here, the intercept coefficient represents the average reading score, which is 500. It shows a large variance (1,594) between countries which is significant using the chi-square test. The reading scale was standardized to a mean of 500 and a standard deviation of 100, across all pupils and countries. Thus the total variance is 10,000, and the between-country variance of 1,594 is about 16.0% of the total variance. The second model includes the methodological variables mean age and school grade. Neither has a significant effect, and the residual variance is still large (1,567) and significant. The residual variance has dropped by 30, which means that about 1.9% of the original between-country variance has been explained. Clearly, the small differences in age and grade do *not* explain the differences between countries.

Table 21.4. Meta-analysis Models for Reading Data

| Model: | Intercept only | | | Methodological vars. | | | + Conceptual vars. | | |
|------------------------|----------------|--------|----------|----------------------|---------|----------|--------------------|---------|----------|
| | coef. | (s.e.) | <i>p</i> | coef. | (s.e.) | <i>p</i> | coef. | (s.e.) | <i>p</i> |
| Fixed part Predictor | | | | | | | | | |
| Intercept | 500 | (7.7) | 0.00 | 792 | (198.9) | 0.00 | 452 | (170.8) | 0.02 |
| Mean age | | | | -.29 | (22.7) | 0.21 | 2.8 | (19.4) | 0.88 |
| Grade | | | | -.25 | (15.6) | 0.88 | 4.6 | (13.6) | 0.73 |
| Economic | | | | | | | 1.0 | (10.3) | 0.97 |
| Health | | | | | | | 22.0 | (8.4) | 0.02 |
| Literacy | | | | | | | 12.0 | (10.7) | 0.27 |
| Random part | | | | | | | | | |
| Variance | | 1,597 | | | 1,567 | | | 816 | |
| χ^2 (df) <i>p</i> | 4,274 | (26) | 0.00 | 3,604 | (24) | 0.00 | 2,701 | (21) | 0.00 |

The third and final model adds conceptual variables that describe theoretically important differences between the countries: in our example, economic status, average health, and overall literacy level. Only the health indicator shows a significant effect. In the third model, the residual variance is much smaller (816) but still significant. Compared to the empty first model, 48.9% of the variance is now explained. The methodological variables therefore explain a (nonsignificant) 1.9% of the variance, while the conceptual variables describing differences between countries explain a (significant) 47% of the variance.

This analysis is intended as an example only. It illustrates the importance of controlling for methodological differences in the study design in international multi-site studies. Even if cross-national studies are intended to be similar in design, in practice some differences are almost inevitable. Consequently, such differences must be controlled for at the analysis stage.

21.5 CONCLUSION

Meta-analysis is of itself comparative, and it becomes international in many cases simply by including all the relevant publications. In cross-cultural or cross-national comparative meta-analyses, cross-comparisons are the central theoretical issue. This is reflected in both the design and the analysis of such meta-analyses.

Implicit in all meta-analytic comparisons is the assumption that the data are comparable. Thus, the data collection procedures in the individual studies must be similar, and the measures (which often imply translating materials) must be equalized. Bechger, van Schooten, de Glopper, and Hox (1998) introduce the general issues involving such comparisons. In a comparative meta-analysis, the ensuing complications should be anticipated. This means that in the design and coding phase of the meta-analysis, indispensable control variables must be explicitly defined and operationalized. In planned multi-site comparisons, the study design should include measures that improve the comparability of the studies. Elley's (1992, 95–111) careful description of the measures taken in the IEA reading study to increase comparability provides a good example.

In the analysis of comparative meta-analytic data, the first step should be to estimate the between-country variability and its significance. The second step then investigates whether any differences found are attributable to methodological differences in the procedures. The third step investigates explanatory variables at the country (or culture) level. This last step is the most interesting one in cross-cultural or cross-national research. The variance between countries must be decomposed into sampling variance, variance due to methodological differences, and systematic and substantively interesting variance. This requires advanced statistical modeling techniques. A classical approach is weighted regression analysis of study outcomes (Hedges and Olkin 1985; Lipsey and Wilson 2001). A powerful alternative is multilevel meta-analysis. This uses iterative Maximum Likelihood estimation, which is an improvement on weighted regression analysis, and is available in standard multilevel software.

A problem that occurs frequently in cross-national comparisons is the multicollinearity of the predictor variables. For instance, one of the problems in the comparison of countries is the large impact of Gross National Product (GNP). Many country-level explanatory variables, such as educational level, health status, or average income, are related to GNP. Even average temperature correlates with GNP. Since these correlations are typically high, it is not possible to include all these variables in the analysis. This creates a problem of choice, since these variables are to some degree interchangeable. The problem is magnified because in cross-national comparisons, the sample size for study-level variables is the number of countries. This is typically not very high, which poses other restrictions on the number of country-level variables that can be included. The issues involved here are no different from the issues in disentangling the contribution of correlated explanatory variables in ordinary multiple regression analysis. They are more difficult to solve because the small samples frequently limit the analysis. Tabachnick and Fidell (2001) discuss possible solutions in the context of multivariate regression analysis, such as step-down analyses.

In the last thirty years, many articles and books have been published on how to do meta-analysis. The classical statistical treatment of meta-analysis is the monograph by Hedges and Olkin (1985). A good general introduction, including a nontechnical discussion of classical analysis methods and how to implement these in SPSS, is the recent book by Lipsey and Wilson (2001). The handbook edited by Cooper and Hedges (1994) is an excellent reader that gives a thorough overview.

The basic statistics for meta-analyses can be programmed easily or calculated by hand. One dedicated program for meta-analysis is by Ralf Schwarzer (Schwarzer 1989). It is user-supported software, meaning that any user may copy and distribute it as long as no charge is made. Lipsey and Wilson (2001) describe procedures for meta-analysis using SPSS, including a method for weighted least squares regression analysis. However, for modeling strongly heterogeneous data, multilevel meta-analysis is superior, because it estimates the between-studies variance using maximum likelihood methods instead of a simple plug-in estimate. Multilevel meta-analysis can be carried out with any multilevel software that supports imposing a constraint on the lowest level variance component. This is possible in the generally available dedicated multilevel software HLM (Raudenbush et al. 2000) and MLwiN (Rasbash 2000) and in the mixed model procedure in the general package SAS (Littell et al. 1996). For details, see Hox (2002) or Hox and de Leeuw (2001).