Marketing Letters 12:2, 157–169, 2001

Meta-Analysis in Marketing when Studies Contain Multiple Measurements

TAMMO H.A. BIJMOLT AND RIK G.M. PIETERS
Department of Marketing, Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands
E-mail: t.h.a.bijmolt@kub.nl

Received July 1999; Revised July 2000; Accepted September 2000

Abstract

Most meta-analyses in marketing contain studies which themselves contain multiple measurements of the focal effect. This paper compares alternative procedures to deal with multiple measurements through the analysis of synthetic data sets in a Monte Carlo study and a re-analysis of a published marketing data set. We show that the choice of procedure to deal with multiple measurements is by no means trivial and that it has implications for the results and for the validity of the generalizations derived from meta-analyses. Procedures that use the complete set of measurements outperform procedures that represent each study by a single value. The commonly used method of treating all measurements as independent performs reasonably well but is not preferable. We show that the optimal procedure to account for multiple measurements in meta-analysis explicitly deals with the nested error structure, i.e., at the measurement level and at the study level, which has not been practiced before in marketing meta-analyses.

Key words: generalizations, meta-analysis, multiple measurements, Monte Carlo simulation

1. Introduction

Meta-analysis integrates the findings of separate studies to determine the overall size of an effect, and to determine the impact of moderating variables on the effect size. It is an important methodological tool to generate generalizable statements about marketing phenomena, and eventually to derive laws of marketing. To fulfill its promise, the meta-analytic procedures need to be reliable and valid, i.e., they should be able to detect the true effect size and the true impact of moderator variables.

Most meta-analyses in marketing contain studies which themselves contain multiple measurements of the effect under study. The distribution of the number of measurements per study is usually highly skewed. For example, the meta-analysis on price elasticity by Tellis (1988) examined 42 studies, which contained a total of 367 measurements with a maximum of 103 in a single study. The average number of measurements of the effect size per study varies considerably, with a low average of 1.2 for Lynn (1991) and a high average of 14.3 for Churchill et al. (1985). Often, meta-

analysis papers mention but do not justify the choice of the specific procedure to deal with multiple measurements within studies (e.g. Brown, Homer, and Inman 1998; Churchill et al. 1985; Sultan, Farley, and Lehmann 1990). Sometimes it is mentioned that alternative procedures were applied and that the results seemed similar (e.g. Abernethy and Franke 1996; Brown and Stayman 1992; Sheppard, Hartwick, and Warshaw 1988). However, it is not obvious whether all procedures are equally appropriate, and how severe the consequences are of using possibly inappropriate procedures to deal with multiple measurements.

This study examines meta-analytic procedures that are used when one or more of the studies in the sample contain multiple measurements of the effect size. We will demonstrate that the specific procedure selected has a significant impact on the results and hence on the validity of the generalizations from the meta-analysis. We consider two general approaches to deal with multiple measurements: (a) represent each study by a single value and, (b) use the complete set of measurements. Six specific procedures of the approaches are compared on their ability to detect the true effect size in a Monte Carlo study, which allows us to examine the performance of the procedures in a well-conditioned setting, with systematic variation of the relevant design factors. In addition, we compare the procedures on a published meta-analysis in marketing to determine how important the choice of procedures is in a practical sense, i.e., whether it really matters.

2. Meta-analysis with Multiple Measurements within Studies

2.1. General Model Structure

Suppose a meta-analysis is conducted to assess the size of a certain effect and the impact of moderator variables on the effect size. A number of studies (indicated as $s = 1, \ldots, S$) are identified from which measurements of the effect can be obtained. For each study one or more measurements of the effect y_{ms}, $m = 1, \ldots, M_s$, are reported. Let $M = \sum_{s=1}^{S} M_s$ denote the total number of measurements.

The measurements of the effect size y_{ms} are to be explained by scores on $k = 1, \ldots, K$ moderator variables, denoted as $x_{k,ms}$. Measurements within a study share values on several moderator variables, e.g. the year of publication and the type of subjects, whereas the measurements may differ on other moderator variables, e.g. the response scale applied to measure the effect. Hence, moderators at the measurement level as well as at the study level may have an impact on the measurements of the effect y_{ms}. The fact that measurements of the effect are not independent within a study, leads to a nested error structure, in which the error variance is decomposed into error at the measurement level e_{ms} and error at the study level u_s. Measurements within the same study share the error component u_s. Measurements of the effect size are generally modeled as a linear function of the overall effect size and the moderator variables, e.g. in the parameter adjustment approach (Farley and Lehmann 1986; Farley, Lehmann, and Sawyer 1995). This leads

to the following general model for meta-analyses with multiple measurements within studies:

$$y_{ms} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{k,ms} + e_{ms} + u_s,$$

where the error components e_{ms} and u_s are assumed to be normally distributed with zero mean and variances σ_e^2 and σ_u^2, respectively.

Estimates for the constant and slope parameters $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K]$ can be obtained by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{T}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{T}^{-1}\mathbf{Y},$$

where $\mathbf{T} = \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}$ and $\mathbf{W}$ is a diagonal matrix with weights and $\boldsymbol{\Sigma}$ is a block-diagonal matrix with error variances.

Estimates for the error variances $\hat{\boldsymbol{\sigma}}^2 = [\hat{\sigma}_e^2, \hat{\sigma}_u^2]$ can be obtained by:

$$\hat{\boldsymbol{\sigma}}^2 = (\mathbf{Z}'(\mathbf{T}^*)^{-1}\mathbf{Z})^{-1}\,\mathbf{Z}'(\mathbf{T}^*)^{-1}\,\mathrm{vec}(\mathbf{Y}^*),$$

where $\mathbf{Z}$ is a $M^2 \times 2$ design matrix with 1 indicating that a specific element of $\boldsymbol{\Sigma}$ contains $\hat{\sigma}_e^2$ respectively $\hat{\sigma}_u^2$, $\mathbf{Y}^* = \mathrm{E}[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})']$ is the cross-product matrix of residuals, and $\mathbf{T}^*$ is $\mathbf{W}\boldsymbol{\Sigma}\mathbf{W} \otimes \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}$.

The estimation procedure (see e.g. Goldstein 1995; Goldstein and Rasbash 1992) iterates between two steps: one in which $\hat{\boldsymbol{\beta}}$ is obtained given $\hat{\boldsymbol{\sigma}}^2$ and one in which $\hat{\boldsymbol{\sigma}}^2$ is obtained given $\hat{\boldsymbol{\beta}}^2$. The procedure starts with obtaining $\hat{\boldsymbol{\beta}}$ given some starting values for $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$, e.g. $\hat{\sigma}_e^2 = \hat{\sigma}_u^2 = 1$.

2.2. Alternative Approaches to Multiple Measurements

There are two general approaches to deal with multiple measurements within studies. For each approach we distinguish various specific procedures, which can be obtained through alternative specifications for $\mathbf{W}$ and $\boldsymbol{\Sigma}$.

First, in the *single value approach*, each study in the meta-analysis is represented by a single value. This can be the *average measurement* per study (Hunter and Schmidt 1990; Rosenthal 1991; Rosenthal and Rubin 1986), which has been used several times in marketing meta-analyses (e.g. Brown, Homer, and Inman 1998; Brown and Stayman 1992; Sheppard, Hartwick, and Warshaw 1988). The single value could also be the *median measurement* in a study (Rosenthal 1991; Rosenthal and Rubin 1986). This procedure has not been applied in marketing meta-analyses to our knowledge. A stochastic component can be introduced through *random selection* of one measurement per study. This random selection procedure has been applied in a meta-analysis on TV advertising effectiveness by Lodish et al. (1995). In the single value approach, the effect of multiple measurements is

aggregated or sampled out. The values in the resulting data sets are independent and there is no need to weight them, that is $\mathbf{\Sigma} = \sigma_e^2 \mathbf{I_S}$ and $\mathbf{W} = \mathbf{I_S}$, where $\mathbf{I_S}$ is the identity matrix of order $S \times S$. The resulting estimates $\hat{\boldsymbol{\beta}}$ are ordinary least squares estimators.

Second, in the *complete set approach*, all measurements are included individually in the analysis. The most commonly used procedure in marketing meta-analyses is to incorporate the values of all measurements within studies and to treat these measurements as *independent replications* (e.g. Abernethy and Franke 1996; Churchill et al. 1985; Peterson 1994; Szymanski et al. 1993; Tellis 1988; Trappey 1996). For this procedure, all weights equal 1, $\mathbf{W} = \mathbf{I_M}$, where $\mathbf{I_M}$ is the identity matrix of order $M \times M$, and the error at the study level is assumed zero $\mathbf{\Sigma} = \sigma_e^2 \mathbf{I_M}$. The resulting estimates $\hat{\boldsymbol{\beta}}$ are ordinary least squares estimators.

Studies with many measurements may have a larger effect on the results of the meta-analysis than studies with few measurements (Rosenthal 1991, p. 27). To account for this some meta-analyses treat multiple measurements as *independent weighted replications*. Sethuraman (1995) has applied this procedure in marketing. For the procedure that weights the measurements, we introduce the weights $w_{ms} = M/M_s S$. Now each study contributes equally to the meta-analysis since $\sum_{m=1}^{M_s} w_{ms} = M/S$, for all $s = 1, \ldots, S$ and that the sum of the weights equals the total number of measurements since $\sum_{s=1}^{S} \sum_{m=1}^{M_s} w_{ms} = S$. To apply these weights, the diagonal matrix $\mathbf{W}$ is formed by placing the square root of the weights onto the diagonal of an $M \times M$ matrix, with zero's as off-diagonal elements. The error at the study level is assumed zero ($\mathbf{\Sigma} = \sigma_e^2 \mathbf{I_M}$). The resulting estimates $\hat{\boldsymbol{\beta}}$ are weighted least squares estimators.

Finally, multiple measurements within a study may be treated explicitly as *dependent replications* by specifying a nested error structure (Raudenbusch, Becker, and Kalaian 1988). To our knowledge, this procedure has not yet been applied in marketing meta-analyses. For the procedure that treats measurements within a study as dependent replications, an $M \times M$ block-diagonal matrix $\mathbf{\Sigma}$ is constructed as follows. For the error structure at the study level, S matrices of size $M_s \times M_s$ with elements σ_u^2 are placed as blocks on the diagonal of $\mathbf{\Sigma}$. For the error variance at the measurement level, σ_e^2 is added to the diagonal of $\mathbf{\Sigma}$. The matrix $\mathbf{\Sigma}$ is used in the estimation method, while all weights equal 1 ($\mathbf{W} = \mathbf{I_M}$). The resulting estimates $\hat{\boldsymbol{\beta}}$ are iterative generalized least squares estimates (Goldstein 1995; Goldstein and Rasbash 1992). On *a priori* statistical grounds the *dependent replications* procedure is preferable, because it uses all the available information and accounts for the nested structure in the data. But does it really matter in meta-analyses to account for all information and for the error-components? We conduct two studies to address this question.

3. Monte Carlo Comparison

3.1. Study Design and Data Generation

A Monte Carlo study is performed, in which the true values of the overall effect size and the impact of moderator variables are known. Synthetic data sets are generated which

systematically vary on six design factors corresponding to important characteristics of marketing meta-analyses. The data sets are generated using the following procedure. First, the *number of studies* is determined to be 20, 50 or 80. Then the numbers of measurements per study are determined by the *average number of measurements* ($\mu = 2$, 5 or 8) and the *variation in the number of measurements*. The latter factor can take the following levels: no variation (100%: μ), low variation (20%: 1; 60%: μ 20%: $(2 \times \mu) - 1$), and high variation (60%: 1; 20%: μ; 20%: $(4 \times \mu) - 3$). The factor levels of the number of studies, the average number of measurements, and variation in the number of measurements reflect the range of values regularly observed in marketing meta-analyses. Given the number of studies and the number of measurements within each study, we generate values for the moderator variables, where the values for the measurement level moderators vary across all measurements while the values for the study level variables vary only between studies and are equal for measurements within a single study. Values of the measurement level moderator variables are generated as follows: two continuous variables drawn from N(0,1) and two dichotomous variables with equal probability being -1 or 1. This procedure is replicated for values of the study level moderator variables. From these eight moderator variables the true values of the dependent variable, $y_{ms}^{(true)}$, are derived, by adding the moderator variables, half of which are multiplied with $\beta_k = 0.50$ and half with $\beta_k = 0.00$ and the appropriate *constant* β_0 (0.00, 0.25, or 0.50). Since all moderator variables have an expected value of zero, these constants can be interpreted as overall effect sizes or average correlations. The variance of the true values, $y_{ms}^{(true)}$, equals 1.00. Next, error components e_{ms} and u_s are drawn from N(0, σ_e^2) and N(0, σ_u^2), respectively, where the *error variance at the measurement level* is low (0.10), medium (1.00), or high (2.00) and the *error variance at the study level* is low (0.10), medium (1.00), or high (2.00). The low error variance levels correspond to virtually no error at the measurement or the study level. The medium variance levels equal the variance of the true structure, and the high error variance levels are twice the variance of the true structure. Finally, the observed values of the dependent variable, $y_{ms}^{(obs)}$, are generated by adding the corresponding error components e_{ms} and u_s to $y_{ms}^{(true)}$. Each factor in the design has three levels and two replications are used in each cell, resulting in a design of 1458 data sets. These data sets serve as input for the six meta-analysis procedures.[1]

[1]One might argue for an alternative procedure to those examined in this paper, namely: including a dummy variable for each study in the meta-analysis. Since this would dramatically alter the set of moderator variables, we did not include this procedure in our main study. To examine the performance of this procedure, we generated data sets varying only those factors deemed most relevant, namely the number of studies (20, 50, 80) and the average number of measurements per study (2, 5, 8). We used the intermediate levels of all other factors from our main study design and generated 10 replications for each of the nine cells. As a benchmark, we used the independent procedure, which yielded an average correlation between the true and predicted values of 0.91, and effects of both factors highly similar to those in the main study. The study-dummy procedure, however, performed dramatically worse with an average correlation of only 0.27, and no significant effect of the number of studies or the number of measurements were found. In most cases, the parameters estimates for the moderator effects, including the dummy variables, diverged extremely far from the true values. This result can be explained by the fact that the dummy variables introduce strong multicollinearity. Estimation of the study dummies is problematic and takes up a substantial proportion of the degrees of freedom, because many studies contain only one or two measurements. Hence, we advise not to apply this study-dummy procedure.

Table 1. Average Correlations Between True and Predicted Values*

		Meta-analysis Procedure (a)					
		Single Value Approach			Complete Set Approach		
Factors	Levels	Average	Median	Random	Independent	Weighted	Dependent
Number of studies (a)	20	0.71	0.67	0.71	0.83	0.81	0.86
	50	0.86	0.84	0.86	0.92	0.91	0.94
	80	0.91	0.88	0.91	0.95	0.93	0.96
Average number of	2	0.82	0.82	0.83	0.88	0.86	0.90
measurements	5	0.82	0.78	0.82	0.91	0.88	0.92
per study (a, b)	8	0.82	0.80	0.83	0.92	0.90	0.94
Variation in number	None	0.83	0.79	0.83	0.91	0.91	0.92
of measurements	Low	0.82	0.80	0.83	0.91	0.89	0.92
per study (a, b)	High	0.82	0.80	0.82	0.88	0.83	0.92
Error variance at the	Low	0.88	0.85	0.88	0.92	0.90	0.94
measurement level	Medium	0.82	0.80	0.83	0.90	0.88	0.92
(a, b)	High	0.77	0.75	0.77	0.88	0.86	0.90
Error variance at	Low	0.88	0.87	0.88	0.96	0.95	0.96
the study level (a, b)	Medium	0.82	0.79	0.83	0.90	0.88	0.92
	High	0.77	0.73	0.77	0.85	0.82	0.88
Constant	0.00	0.82	0.79	0.83	0.90	0.88	0.92
	0.25	0.82	0.80	0.83	0.90	0.88	0.92
	0.50	0.83	0.80	0.82	0.90	0.88	0.92
Total mean		0.82	0.80	0.83	0.90	0.88	0.92

*a = Main effect significant at $\alpha = 0.0$l; b = Interaction effect with meta-analysis procedure significant at $\alpha = 0.0$l.

3.2. Results

To compare the performance of the procedures, three sets of evaluation criteria are used: (1) overall model recovery, (2) parameter significance testing, (3) parameter estimation accuracy.

3.2.1. Overall Model Recovery. We assess overall model recovery by comparing the predicted values $y_{ms}^{(pred)}$ with the true error-free values $y_{ms}^{(true)}$. The similarity between these values is measured by the product-moment correlation. Repeated measurement ANOVA[2] is performed (after Fisher transformation of the correlations) and for each procedure the average correlation is reported for each design factor (see Table 1).

The *recovery of the true measurements* varies significantly between the meta-analysis procedures. The three single value procedures are clearly outperformed by the procedures that analyze all measurements. The independent replications procedure and the weighted

[2]For reasons of parsimony and interpretability, we restrict all ANOVA's to main effects and interactions including at most two design factors.

independent replications procedure recover the true values somewhat better, but accommodating the dependency between measurements from a single study results in the most accurate recovery.

The number of studies in the meta-analysis has a large positive effect on the recovery. This effect holds for each meta-analysis procedure. The effect of the number of measurements per study, however, varies substantially across procedures. The average number of measurements has a positive effect on the recovery by the procedures using all measurements, whereas it has no effect on the recovery by the average measurement, median measurement, and random selection procedures. Variation in the number of measurements per study has little effect for most procedures, with the exception of a negative effect for the independent replications and weighted independent replications procedures. If the distribution of the number of measurements per study is highly skewed, the independent replications and weighted independent replications procedures recover the true values hardly better than the single value procedures. As expected, an increase in the error level at either the measurement level or the study level has large negative effects on the recovery performance of the various procedures. The effect of increasing levels of error variance at the measurement level is largest for the three single value procedures. Compared to other procedures, model recovery of the dependent replication procedure is least affected by increasing levels of error variances at the measurement or study level.

3.2.2. Parameter Significance Testing. We assess parameter significance testing of the various procedures by evaluating the type I and type II error rates for each procedure. A good estimation procedure should identify moderator variables generated with $\beta_k = 0.50$ as significant and moderator variables generated with $\beta_k = 0.00$ as not significant. Repeated measurement ANOVA's are performed on both sets of p-values, and for each procedure the average type I and type II error rates are reported given the factors in the design and the moderator variables (see Table 2).

Across all conditions, the dependent replications procedure has the lowest *type I error rate*, followed by the three single value procedures, then followed by the independent replications and weighted replications procedures. The type I error rates differ substantially between the various moderator variables. For the measurement level variables, the independent replications and weighted replications procedures perform about equal to the dependent replications procedure, and all three clearly outperform the three single value procedures. However, the type I error rate for the study level variables is extremely high for the independent replications and weighted replications procedure. For most procedures, the design factors do not have a large impact on the p-values of the zero-effect moderator variables ($\beta_k = 0.00$), and consequently not on the type I error rate. All procedures perform better if the number of studies increases.

Considering all moderator variables simultaneously, the *type II error rates* is lower for the independent replications and weighted procedures, than for the dependent procedure. The three single value procedures perform much worse. The dependent replications procedure outperforms the independent replications and weighted replications procedures with respect to the measurement level variables, but it is outperformed by these procedures with respect to the study level variables. Three design factors have a considerable impact

Table 2. Type I and Type II Error Rates of Significance Tests*

		Meta-analysis Procedure											
		Single Value Approach						Complete Set Approach					
Factors	Levels	Average		Median		Random		Independent		Weighted		Dependent	
Results aggregated across the design factors:													
Constant = 0.00		0.12		0.12		0.10		0.30		0.34		0.09	
Constant = 0.25			0.64		0.56		0.66		0.34		0.33		0.54
Constant = 0.50			0.30		0.23		0.32		0.14		0.15		0.24
Measurement-level moderators		0.11	0.33	0.11	0.49	0.10	0.31	0.06	0.07	0.06	0.07	0.07	0.05
Study-level moderators		0.11	0.32	0.12	0.26	0.11	0.32	0.30	0.13	0.35	0.14	0.09	0.24
Results aggregated across the moderator variables:													
Number of studies	20	0.18	0.52	0.20	0.53	0.16	0.50	0.20	0.21	0.23	0.21	0.10	0.26
	50	0.09	0.30	0.09	0.35	0.08	0.29	0.18	0.06	0.19	0.07	0.06	0.10
	80	0.07	0.17	0.07	0.25	0.07	0.17	0.17	0.02	0.19	0.03	0.06	0.05
Average number of measurements per study	2	0.11	0.33	0.12	0.35	0.10	0.33	0.12	0.19	0.15	0.20	0.08	0.20
	5	0.12	0.33	0.13	0.39	0.11	0.31	0.20	0.07	0.22	0.08	0.08	0.12
	8	0.11	0.33	0.11	0.39	0.10	0.31	0.23	0.04	0.25	0.05	0.06	0.10
Variation in number of measurements per study	None	0.11	0.32	0.12	0.38	0.10	0.29	0.16	0.09	0.16	0.09	0.07	0.13
	Low	0.12	0.34	0.11	0.39	0.10	0.33	0.17	0.10	0.19	0.11	0.07	0.13
	High	0.11	0.33	0.12	0.36	0.10	0.33	0.22	0.11	0.28	0.13	0.08	0.16
Error variance at the measurement level	Low	0.12	0.18	0.12	0.23	0.11	0.17	0.24	0.05	0.27	0.05	0.08	0.08
	Medium	0.11	0.35	0.11	0.39	0.09	0.33	0.16	0.10	0.19	0.11	0.06	0.15
	High	0.11	0.46	0.11	0.50	0.11	0.46	0.15	0.15	0.16	0.16	0.08	0.19
Error variance at the study level	Low	0.12	0.19	0.13	0.20	0.11	0.18	0.13	0.05	0.14	0.06	0.07	0.05
	Medium	0.10	0.34	0.11	0.40	0.10	0.32	0.19	0.10	0.21	0.10	0.08	0.14
	High	0.11	0.46	0.12	0.52	0.11	0.45	0.23	0.15	0.27	0.16	0.07	0.23
Constant	0.00	0.11	0.34	0.12	0.38	0.11	0.32	0.18	0.10	0.21	0.10	0.07	0.14
	0.25	0.11	0.33	0.11	0.38	0.10	0.31	0.19	0.10	0.21	0.11	0.08	0.14
	0.50	0.12	0.32	0.12	0.36	0.10	0.32	0.18	0.09	0.20	0.11	0.07	0.14
Total mean		0.11	0.33	0.12	0.37	0.10	0.32	0.18	0.10	0.21	0.11	0.07	0.14

*Cells contain type I and type II error rates, respectively.

on the p-values of moderator variables with non-zero effects ($\beta_k = 0.50$): the type II error rate decreases with an increase of the number of studies and a decrease of the error variance at either the measurement or the study level. The effect of the average number of measurements per study differs substantially between procedures. For the procedures using all measurements, an increase of the number of measurements results in lower type II error rates, for the average measurement and the random selection procedures this number has little to no effect, whereas for the median measurement procedure an increase of the average number of measurements results in higher type II error rates.

3.2.3. Parameter Estimation Accuracy. The parameter estimates of the moderator effects should be close to the true values of $\beta_k = 0.00$ and $\beta_k = 0.50$. Bias is measured through

the estimation error ($\beta^{(est)} - \beta^{(true)}$). This reflects the tendency of systematically under- or overestimating the true values. (In-)consistency is measured through the absolute estimation error $|\beta^{(est)} - \beta^{(true)}|$. This reflects the tendency of deviating from the true values, irrespective of underestimation or overestimation. Repeated measurement ANOVA's are performed and for each procedure the mean estimation errors and the mean absolute errors are reported for each moderator variable and each design factor (see Table 3).

Across all variables and design factors, each of the procedures has a slight tendency to underestimate the true parameter values. The average *bias of the parameter estimates* is smallest and in a practical sense negligible for the dependent replications procedure, somewhat larger for the independent and weighted procedures, and much larger for the single value procedures. Clearly, the median measurement procedure has by far the largest

Table 3. Bias and Consistency of Parameter Values

		Meta-analysis Procedure											
		Single Value Approach						Complete Set Approach					
Factors	Levels	Average		Median		Random		Independent		Weighted		Dependent	
Results aggregated across the design factors:													
Constant, $\beta_0 = 0.00$		0.02	0.22	0.06	0.19	0.06	0.22	−0.07	0.18	−0.16	0.21	−0.03	0.16
Constant, $\beta_0 = 0.25$		0.01	0.21	0.13	0.19	−0.05	0.21	0.16	0.18	0.22	0.21	0.07	0.16
Constant, $\beta_0 = 0.50$		−0.13	0.22	0.06	0.19	0.05	0.21	−0.01	0.18	0.01	0.21	−0.02	0.16
Measurement-level moderators		−0.11	0.21	−0.75	0.23	−0.06	0.21	−0.04	0.09	−0.03	0.09	−0.01	0.07
Study-level moderators		0.03	0.21	0.00	0.19	0.01	0.21	−0.01	0.17	−0.01	0.21	0.01	0.16
Results aggregated across the moderator variables:													
Number of studies	20	−0.11	0.33	−0.45	0.32	−0.04	0.32	−0.06	0.19	−0.07	0.22	−0.04	0.17
	50	0.03	0.17	−0.30	0.17	−0.01	0.17	0.05	0.11	0.06	0.13	0.04	0.10
	80	−0.05	0.13	−0.40	0.14	−0.05	0.13	−0.05	0.09	−0.04	0.10	−0.03	0.07
Average number of measurements per study	2	−0.02	0.21	−0.16	0.21	−0.03	0.21	−0.03	0.15	−0.03	0.17	−0.01	0.14
	5	−0.05	0.21	−0.55	0.21	−0.04	0.21	−0.05	0.13	−0.05	0.15	−0.03	0.11
	8	−0.06	0.21	−0.44	0.21	−0.03	0.20	0.02	0.12	0.03	0.13	0.01	0.09
Variation in number of measurements per study	None	−0.03	0.21	−0.45	0.22	−0.04	0.21	−0.00	0.12	−0.00	0.12	0.01	0.11
	Low	−0.05	0.21	−0.40	0.21	−0.04	0.20	−0.04	0.13	−0.04	0.13	−0.02	0.11
	High	−0.05	0.21	−0.30	0.21	−0.02	0.21	−0.02	0.15	−0.01	0.19	−0.01	0.12
Error variance at the measurement level	Low	0.05	0.15	−0.13	0.17	0.08	0.15	0.07	0.11	0.07	0.13	0.06	0.08
	Medium	−0.16	0.21	−0.45	0.21	−0.12	0.21	−0.10	0.13	−0.11	0.15	−0.08	0.12
	High	−0.02	0.27	−0.57	0.25	−0.06	0.25	−0.03	0.15	−0.01	0.17	−0.01	0.14
Error variance at the study level	Low	−0.06	0.15	−0.37	0.14	−0.04	0.15	−0.02	0.08	−0.02	0.09	−0.02	0.08
	Medium	−0.10	0.21	−0.42	0.21	−0.00	0.21	−0.04	0.14	−0.03	0.15	−0.01	0.12
	High	−0.03	0.26	−0.35	0.28	−0.06	0.26	0.00	0.18	−0.00	0.20	0.00	0.14
Constant	0.00	−0.07	0.21	−0.41	0.21	−0.03	0.20	−0.02	0.13	0.00	0.15	−0.01	0.11
	0.25	0.02	0.21	−0.40	0.21	−0.01	0.21	0.01	0.13	0.03	0.15	0.02	0.11
	0.50	−0.07	0.22	−0.35	0.22	−0.08	0.21	−0.05	0.13	−0.08	0.14	−0.03	0.11
Total mean		−0.04	0.21	−0.38	0.21	−0.03	0.21	−0.02	0.13	−0.02	0.15	−0.01	0.11

*Cells contain bias (mean estimation error $\times 10^{-1}$) and consistency (mean absolute error), respectively.

negative bias in the parameter estimates, which is largely due to a systematic and serious underestimation of non-zero measurement level effects. An increase in the number of measurements per study leads to a larger bias for the median measurement procedure. The significant effect of the error variance at the measurement level is due to the case of low error variance, when most procedures tend to overestimate parameter values, especially those with a true value of 0.00. However, in general the design factors have hardly any influence on the bias of the parameter estimates.

Considering all data sets and variables, the dependent replications procedure results in the highest *consistency of the parameter estimates*, followed by the independent replications procedure, the weighted replications procedure, and the three single procedures. The procedures using all measurements reach a higher consistency for measurement level variables than for study level variables. For each of the procedures, the parameter estimates deviate more from the true values in case of a small number of studies or high error variance level at the measurement or the study level. These effects are especially pronounced for the three single value procedures. Next, if the average number of measurements per study increases, the consistency of the estimates from the independent, weighted, and dependent replications procedures increases. For the single value procedures, this design factor does not have a positive effect. On the other hand, if the variation in the number of measurements per study increases, the consistency of the estimates by the independent and weighted procedures decreases. For the other four procedures this design factor does not have a substantial effect.

In the Monte Carlo study we examined synthetic data sets. To examine how the various procedures perform in a real-life context, we re-analyzed a published meta-analysis in marketing.

4. Empirical Comparison: Intention-behavior Consistency

Sheppard, Hartwick, and Warshaw (1988) performed an important meta-analysis on the theory of reasoned action to examine the correlation between measures of behavioral intention and behavior. The meta-analysis included 31 papers, containing 34 separate studies with an average of 2.6 measurements of the correlation. Most papers contributed a single effect size measurement. However, one paper contained two studies that contributed 18 measurements each, and another paper with a single study contributed 11 measurements. These figures are representative for meta-analyses in marketing. Since the average number of measurements per study is substantially larger than one and the distribution is highly skewed across studies, the procedure to handle multiple measurements in this meta-analysis may potentially affect its results.

The observed correlations vary between 0.10 and 0.96. Sheppard et al. (1988) examined the following measurement level moderating variables: the type of activity or outcome (ACTIVITY; 0 = behavior, 1 = goal), whether or not the behavior measure involves a choice among alternative activities or outcomes (CHOICE; 0 = no choice, 1 = choice), the type of intention measure (MEASURE; 0 = intention or not specified, 1 = estimate). Following Kayande and Bhargava (1994), we added the year of publication (YEAR; 0 = 1970–1983, 1 = 1984–1986) as a potential moderator that varies at the study level.

We estimated a linear model for each of the procedures to deal with multiple measurements, using the Fisher transformed correlation as the dependent variable and the moderators aforementioned as explanatory variables (see Table 4).

The constant is highly significant in each of the procedures, but the estimated value varies considerably. Furthermore, the procedures differ considerably with respect to the significance level of the moderator effects. Across all moderators, the three complete set procedures yield more significant effects ($\alpha = 0.05$) than the three single value procedures. The procedure treating the measurements as independent yields the highest number of significant effects, namely three. The procedure selecting the median measurement generally has the highest p-values. If $\alpha = 0.10$ is used, still none of the moderator effects is significant for the median measurement procedure, whereas the number of significant effects for the other procedures varies between two and four.

The signs of the main effect parameters tend to be the same across procedures. Only the sign of the parameter estimate of MEASURE varies across procedures, but the effect does not differ significantly from zero for any of the procedures. However, the parameter estimates of the interactions vary considerably between the meta-analytic procedures, from -0.05 to 0.51 for ACTIVITY $\times$ MEASURE and from -0.01 to 0.19 for CHOICE $\times$ MEASURE. These interactive effects are not significant at $\alpha = 0.05$ for any procedure. For the independent, average measurement, and random selection procedures, however, the ACTIVITY $\times$ MEASURE interaction is positive and significant at $\alpha = 0.10$.

5. Conclusion and Discussion

We compared alternative procedures to deal with multiple measurements through the analysis of synthetic data sets in a Monte Carlo study and an actual marketing

Table 4. Meta-analysis of the Intention-behavior Correlation: Results of Alternative Procedures

	Meta-analysis Procedure*					
	Single Value Approach			Complete Set Approach		
	Average	Median	Random	Independent	Weighted	Dependent
Constant	0.82 (< 0.01)	0.71 (< 0.01)	0.83 (< 0.01)	0.75 (< 0.01)	0.69 (< 0.01)	0.76 (< 0.01)
ACTIVITY	−0.30 (0.07)	−0.22 (0.18)	−0.36 (0.04)	−0.28 (0.01)	−0.18 (0.22)	−0.24 (0.05)
CHOICE	0.15 (0.36)	0.23 (0.16)	0.16 (0.36)	0.27 (0.03)	0.37 (< 0.01)	0.22 (0.13)
MEASURE	−0.13 (0.53)	−0.02 (0.92)	−0.15 (0.49)	0.02 (0.80)	0.08 (0.32)	0.05 (0.76)
ACTIVITY × MEASURE	0.51 (0.07)	0.40 (0.15)	0.50 (0.06)	0.27 (0.09)	−0.05 (0.80)	0.16 (0.35)
CHOICE × MEASURE	0.19 (0.47)	−0.01 (0.98)	0.10 (0.71)	0.00 (0.98)	0.03 (0.85)	0.01 (0.95)
YEAR	−0.30 (0.06)	−0.26 (0.11)	−0.27 (0.10)	−0.20 (0.02)	−0.14 (0.09)	−0.27 (0.05)

*Cells contain the parameter estimate followed by the p-value between brackets.

meta-analysis. The results consistently demonstrate that the specific procedure selected affects the results of the meta-analysis substantially and hence the validity of the empirical generalizations that are based on the results.

The performance of meta-analytic procedures that reduce each study to a single value is generally unsatisfactory. They do not perform very well with respect to recovering the true measurements of the effects. The estimated effects of moderator variables deviate substantially from the true values. Furthermore, they have a low power (around 0.65 on average): moderator variables that are known to have an effect are frequently judged to have no effect. This holds in particular for the median measurement procedure. Because procedures representing each study by a single value result in a serious loss of information, they should in general be avoided in meta-analyses. This recommendation is less obvious than it may seem at first sight: leading meta-analytic theorists such as Hunter and Schmidt (1990) and Rosenthal (1991) recommend to have each study contribute only a single value, and it is common practice to do so in marketing meta-analyses.

Procedures that use the complete set of measurements from each study outperform procedures that represent each study by a single value only. Including all measurements while not accounting for the dependency between measurements from the same study, however, inflates the performance of the independent and weighted procedure. Especially conclusions about study level moderators may be biased. The straightforward procedure of treating all measurements as independent, which is the most commonly applied procedure in marketing meta-analyses, turns out to be a sensible but sub-optimal procedure. Across the board the dependent replications procedure outperforms the independent and weighted procedures, as it recovers the true measurements better, results in lower type I errors rates, a higher power for measurement level moderators, and a smaller bias and higher consistency of the parameter estimates. In general, this should be the procedure of choice in meta-analyses with multiple measurements within studies. To our knowledge the dependent replications procedure has not yet been applied in meta-analyses in marketing. However, as computer programs that allow such dependent replications analysis, e.g. HLM or MLWin, become more widely available this should change in the near future.

The goal of meta-analysis in marketing is to accumulate and generalize results across studies in order to identify the current state of knowledge on a certain substantive matter, and to identify areas for future research from there (Farley and Lehmann 1986; Farley, Lehmann, and Sawyer 1995). Using the proper meta-analysis procedure is paramount since it provides the unbiased view on the state of current marketing knowledge that we are searching for, and points future research into the right direction.

References

Abernethy, Avery M, and George R. Franke. (1996). "The information content of advertising: A meta-analysis," *Journal of Advertising*, 25 (Summer), 1–17.

Brown, Steven P, Pamela M. Homer, and J. Jeffrey Inman. (1998). "A meta-analysis of relationships between ad-evoked feelings and advertising responses," *Journal of Marketing Research*, 35 (February), 114–126.

Brown, Steven P, and Douglas M. Stayman. (1992). "Antecedents and consequences of attitudes toward the ad: a meta-analysis," *Journal of Consumer Research*, 19 (June), 34–51.

Churchill, Gilbert A, Neil M. Ford, Steven W. Hartley, and Orville C. Walker. (1985). "The determinants of salesperson performance: a meta-analysis," *Journal of Marketing Research*, 22 (May), 103–118.

Farley, John U, and Donald R. Lehmann. (1986). *Meta-analyses in marketing: generalizations of response models*. Lexington: Lexington books.

Farley, John U, Donald R. Lehmann, and Alan Sawyer. (1995). "Empirical marketing generalization using meta-analysis," *Marketing Science*, 14(3, part 2), 36–46.

Goldstein, Harvey. (1995). *Multilevel statistical models, second edition*. London: Arnold.

Goldstein, Harvey, and Jon Rasbash. (1992). "Efficient procedures for estimation of parameters in multilevel models based on iterative generalized least squares," *Computational Statistics and Data Analysis*, 13, 63–71.

Hunter, John E, and Frank L. Schmidt. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park: Sage.

Kalaian, Hripsime A, and Stephen W. Raudenbusch. (1996). "A multivariate mixed linear model for meta-analysis," *Psychological Methods*, 1(3), 227–235.

Kayande, Ujwal, and Mukesh Bhargava. (1994). "An examination of temporal patterns in meta-analysis," *Marketing Letters*, 5(2), 141–151.

Lodish, Leonard M, Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Rihardson, and Mary Ellen Stevens. (1995). "How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments," *Journal of Marketing Research*, 32 (May), 125–139.

Lynn, Michael. (1991). "Scarcity effects on value: a quantitative review of the commodity theory literature," *Psychology & Marketing*, 8 (Spring), 43–57.

Peterson, Robert A. (1994). "A meta-analysis of Cronbach's coefficient alpha," *Journal of Consumer Research*, 21 (September), 381–391.

Raudenbusch, Stephen W, Betsy J. Becker, and Hripsime A. Kalaian. (1988). "Modeling multivariate effect sizes," *Psychological Bulletin*, 103, 111–120.

Rosenthal, Robert. (1991). *Meta-analytic procedures for social research*, revised edition. Newbury Park: Sage.

Rosenthal, Robert, and Donald B. Rubin. (1986). "Meta-analytic procedures for combining studies with multiple effect sizes," *Psychological Bulletin*, 99(3), 400–406.

Sethuraman, Raj. (1995). "A meta-analysis of national brand and store brand cross-promotional price elasticities," *Marketing Letters*, 6(4), 275–286.

Sheppard, Blair H, Jon Hartwick, and Paul R. Warshaw. (1988). "The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research," *Journal of Consumer Research*, 15 (December), 325–343.

Sultan, Fareena, John U. Farley, and Donald R. Lehmann. (1990). "A meta-analysis of applications of diffusion models," *Journal of Marketing Research*, 27 (February), 70–77.

Szymanski, David M, Sundar G. Bharadwaj, and P. Rajan Varadarajan. (1993). "An analysis of the market share-profitability relationship," *Journal of Marketing*, 57 (July), 1–18.

Tellis, Gerard J. (1988). "The price elasticity of selective demand: a meta-analysis of econometric models of sales," *Journal of Marketing Research*, 25 (November), 331–341.

Trappey, Charles. (1996). "A meta-analysis of consumer choice and subliminal advertising," *Psychology & Marketing*, 13 (August), 517–530.