Resistance Training:

Identifying Best Practices?

Ross R. Vickers Jr.
Amanda C. Barnard
Linda K. Hervig

Naval Health Research Center

Report No. 10-21

.

Naval Health Research Center
140 Sylvester Rd.
San Diego, California 92106-3521

Resistance Training: Identifying Best Practices?

Ross R. Vickers Jr.
Linda K. Hervig
Amanda Barnard

Warfighter Performance Department
Naval Health Research Center
140 Sylvester Road
San Diego, CA 92106-3521

e-mail: ross.vickers@med.navy.mil

This represents report 10-21 and was supported by the U.S. Department of Defense under work unit 60704.

Abstract

Resistance training increases muscle strength. Muscle strength gains are influenced by program design. This review attempted to identify design choices that would be best practices. A best practice is a design option that produces significantly better results than any other option. To ensure sensitive assessments of program design effects, statistical procedures adjusted for differences in program length, and allowed for the repeated measures structure of the study designs. Untrained individuals benefitted much more from training than trained individuals. Gender had little effect. Age effects differed for men and women. Given the impact of participant characteristics on the training response, the effects of different program design facets were examined separately for programs with untrained and trained participants. Periodization, number of sessions per week, number of sets per session, and intensity (number of repetitions per set) were significant moderators for untrained participants; sets per session and intensity were significant moderators for trained participants. However, comparisons generally showed that no single design option was significantly better than all others. The available evidence may rule out some design choices (e.g., a single set per session), but it is too limited to identify best practices.

Introduction

Resistance training literature reviews have shown definitively that training increases muscle strength (Falk & Tenenbaum, 1996; Payne, Morrow, Johnson, & Dalton, 1997; Peterson, Rhea, & Alvar, 2004, 2005; Rhea & Alderman, 2004; Rhea, Alvar, & Burkett, 2002; Rhea, Alvar, Burkette, & Ball, 2003; Wolfe, LeMura, & Cole, 2004). The same reviews have established that strength gains are influenced by training program designs.

With the effectiveness of resistance training well established, attention shifts to a different question: Is it possible to identify best practices for resistance training? A best practice is a specific program design option that is superior to all other possible choices for that program design facet. For example, the number of sets in each training session is a program design facet. If the cumulative research record indicated that three sets per session produced significantly better results than any other choice for this facet, then three sets per session would be a best practice.

This review attempted to identify best practices based on the available evidence. Two differences from past reviews were introduced to maximize sensitivity to the presence of best practices if they exist. First, logic and common experience lead to the expectation that longer training programs will produce greater training effects. In previous meta-analyses program length has been treated as a categorical variable (e.g., 6-16 weeks vs. 17-40 weeks) rather than a continuous variable (Rhea & Alderman, 2004; Payne et al., 1997; Wolfe et al., 2004). Important information may have been lost by collapsing program length into categories. This study revisited the question, “Do longer programs produce greater training effects?”—with program length treated as a continuous variable. An affirmative answer to this question would raise the secondary question “Has the failure to control for differences in program length distorted the relationships of program design facets to program effectiveness?” This review introduced statistical controls for differences in program length to answer this question.

The second difference between this review and prior reviews involved the treatment of statistical issues. One set of issues derived from the repeated measures structure of the evidence. Resistance training studies often employ multiple strength tests to assess training effects. Each test is an attempt to measure the training program effects. If the training program produces a single common effect for all muscle groups, the use of multiple tests constitutes repeated measurement of that effect. Steps must be taken to deal with the fact that the results for different tests are not independent (Gleser & Olkin, 1994).

A set of related statistical issues arose from the typical design of resistance training studies. Resistance training studies routinely employ repeated measures research designs. Strength tests are administered before the training program begins and again after the program has been completed. The difference between the pre- and post-training scores is the basis for estimating the effect size (ES) for the training program. Steps must be taken to allow for this research design when estimating the ES for a study (Morris & DeShon, 2002).

A third statistical issue derived directly from the current interest in identifying best practices. It is not enough to demonstrate that program design choices effect the size of the training response. Analysis of variance (ANOVA) tests have been used to test the hypothesis that all program design choices have produced equal effects. Rejecting this null hypothesis has only indicated that some options differ from other options. It is not enough to know that differences between options exist. The existence of differences does not guarantee the existence of a best choice. For example, Rhea et al. (2003) showed that the number of sets per session affected the magnitude of the training effect for trained individuals. Four sets per session was identified as the optimal choice because programs embodying that option had produced that largest average effect, ES = 1.17. However, the ES for 4 sets per session was only trivially larger than the average ES for 5 sets per session**,** ES = 1.15. The hypothesis that 4 sets per session produced a stronger effect than 5 sets per session would be rejected unless the sample size was very large. This example has illustrated the point that a significant ANOVA must be followed by analyses that evaluate differences between specific program design options. This review employed post hoc comparisons to determine whether the design option that produced the largest ES was truly a best practice.

This review attempted to identify best practices for several design facets of resistance training programs. Statistical methods were introduced to deal with the program length, the repeated measures structure of the data, and the need for post hoc comparisons to determine whether a significant moderator effect truly identifies a best practice. No other review to date has dealt with all of these issues or employed a formal definition of a best practice. As a consequence, this review provided a different perspective on the available evidence. This review attempted to formally identify best for the program design facets of periodization, number of training sessions per week, number of sets per session, and number of repetitions per set. The number of repetitions per set is a proxy measure for the intensity of the training program. The search for best practices also considered age, gender, and training status as demographic variables that might influence the impact of different program design choices. The concern was that best practices might depend on the type of person being trained.

Methods

Literature Search Procedures

The literature search began by identifying articles that contributed data to previous resistance training meta-analyses (Falk & Tenenbaum, 1996; Payne et al., 1997; Peterson et al., 2005; Rhea & Alderman, 2004; Rhea et al., 2002; Rhea et al., 2003; Wolfe et al., 2004). Subsequent steps centered on a search of the PubMed database. The search terms “resistance training or weight training and strength” produced a list of 2,432 candidate articles.

The candidate articles were separated into two groups for further review. The first group consisted of 1,366 articles published between January 1, 2000 and May 16, 2007, the time of the search. The PubMed abstract for each of these articles was examined to determine whether it met the inclusion criteria for this review. Articles were dropped at this point in the search only if the information in the abstract clearly indicated that the study failed to meet at least one of the criteria.

The second group of articles consisted of 1,066 articles published before 2000. The titles of these articles were reviewed. The abstract of the article was reviewed only if the title suggested that the study included an experimental evaluation of one or more resistance training programs, and the article had not been included in any of the meta-analyses cited in the introduction to this paper. The review procedures for this second group of articles were less intensive than those employed for the first group articles. The procedures were relaxed because it was assumed that prior meta-analyses had identified most of the relevant studies conducted prior to 2000.

When the abstract of a study was reviewed, the inclusion criteria were:

1. At least one group in the study had to participate in a resistance training program. In addition to the usual resistance training program studies, this criterion resulted in the inclusion of placebo control groups from studies that evaluated the effects of supplements. In such cases the placebo group underwent training without any additional experimental manipulations. Thus, the effects of resistance training were not confounded with supplement effects and could legitimately be included in an overall evaluation of resistance training programs.
2. Strength measurements had been made prior to, and after, the program. The specific measurements were not a concern at this point in the search.
3. Study participants were healthy. This requirement excluded studies of specific disease populations, including chronic obstructive pulmonary disease, HIV infection, chronic heart failure, diabetes, fibromyalgia, and so forth. The general rule was that a study was excluded if the authors characterized the study population as “patients.” Studies of people who were characterized as obese, hypertensive, or frail were excluded. However, studies of “overweight” individuals were accepted, so weight considerations eliminated only studies of individuals toward the upper end of the excess weight range. The objective in making these exclusions was to eliminate studies that might produce atypical effects because of limitations on the ability to perform training exercises, and/or that involved disease and metabolic processes that might modify the training response.
4. The average study participant had to be at least 16 years of age. This criterion attempted to minimize the confounding of training effects with the effects of normal developmental processes.
5. The study employed isotonic or isoinertial strength measures. Studies that relied on isokinetic or isometric measures (k = 174) were dropped to ensure that the operational definition of strength was the same in each study. This step eliminated measurement methods as a potential source of variation in ES. This criterion was introduced because Payne et al. (1997) have demonstrated that measurement modality affected ES.

The full text of 438 articles that passed the screening procedures was examined to determine whether the studies reported the basic data required for this review. The minimal requirement for retention during this phase of the search was that the study had to provide the statistics required to compute the ES for isoinertial/isotonic strength tests. Specifically:

1. The unit of measurement was pounds or kilograms. Studies that relied on other units of measurement, such as Newtons, pneumatic measures, or percent change, were excluded.
2. The study reported pre- and post-training measures of strength and the standard deviations for those measures. This information was the minimum required to compute ES when combined with assumptions about the magnitude of the pretest–posttest correlation (see Appendix A). Strength was measured more than twice in some studies. When this was the case, ES was computed using the initial and final measurements. Computing the effect for each phase of the training programs would have increased the complexity of the repeated measures problem. Therefore, ES always represented the final cumulative impact of training.

Supplementary searches were conducted because the initial search and data coding took long enough to allow further studies to enter the literature. Also, informal reading of the journals that provided most of the studies identified in the primary search suggested that some resistance training studies might have been missed because the studies were outside the scope of the initial search terms.

The first supplementary search addressed the problem by adding the terms “Training” and “1-RM.” The search term 1-RM was used to identify studies that involved 1-repetition maximum (1-RM) strength measures. This search identified 242 articles, 139 of which were reviewed in detail.

The second supplementary search repeated the initial search, but covered a different time frame. The same keywords were used, but the search was limited to articles published between January, 2007 and February, 2009. This search identified 706 articles. Examination of the article abstracts reduced the number of articles for review to 117 for direct inspection. The inclusion rules for the initial review were employed for these later reviews.

The final database consisted of information from 196 studies that met the review criteria. Control groups from those studies were excluded from the review. With this restriction, 302 samples provided sufficient data to be included in this review. The cumulative sample size was 4,574 study participants.

Table 1

Sample Characteristics

	k	ΣN	Mean	SD	Minimum	Maximum
Age	273	4054	33.20	19.22	15.9	82.0
Height	234	3539	174.19	7.57	154.0	189.0
Weight	270	3941	77.38	11.64	55.0	125.0
Percent body fat	92	1225	21.35	7.37	11.6	39.0
Fat-free mass	58	753	60.18	12.48	27.8	87.0

Note. Statistics describe samples in the analysis, not individuals. The data were not weighted for the computations that generated these descriptive statistics. The statistics describe the population of study samples rather than a population of individuals.

Demographic Characteristics

Height, weight, percent body fat, fat free mass. Height, weight, percent body fat, and fat free mass were coded from descriptive statistics reported in the studies (see Table 1). Note that these statistics describe samples not individuals.

Gender. The samples in most studies consisted entirely of men or entirely of women. Other studies combined the data for men and women when reporting the study findings. A few studies provided no definite information regarding gender. Given this variability, the gender of each sample was coded as male, female, men and women combined, or indeterminate.

Age. Age most often was reported by giving the average ages separately for each treatment group in the study. In other cases, the study reported only the average age for all of the study participants. An age range (e.g., 21 to 32 years) was another common reporting method. Finally, some studies did not report age directly, but provided age-related demographic information (e.g., college students, health community-dwelling elders). Given this variable content, age was reduced to a dichotomy identifying older and younger samples. Where quantitative information was available, samples with an average age ≤50 were classified as younger. Qualitative data were coded based on judgments of the age range that would be typical of the group described. This dichotomy provided a variable that minimized missing data for age because it could be applied to as many samples as possible given the available information.

Training status. Training status was inferred from descriptions of the recent training status of study participants. The initial coding of training status employed five categories: sedentary, recreationally active, athletes in sports that do not routinely involve resistance training, recreational weight trainers, and competitive weight trainers. The last category included athletes who participated in weight lifting competitions and athletes in sports such as football, where weight training is employed to gain a competitive advantage.

Preliminary analyses showed that a dichotomy captured most of the ES variation across the initial five categories. Trained samples consisted of individuals in ongoing resistance training programs. Untrained samples consisted of individuals who either had no prior resistance training experience or who had not trained for at least several months prior to starting the study. The two training status categories are referred to as “trained” and “untrained” participants in the remainder of this paper.

Program Design Facets

Length. Program length was the number of weeks that the training program lasted. In some cases, baseline strength measurements were taken well before beginning the actual training process. In other cases, several sessions with light weights were used to familiarize the study participants with the resistance exercises. In either case, length included only the actual training period as designated by the study’s authors.

Sessions per week. This variable represented the number of times that each exercise in the program was performed during a given week. This number could be less than the total number of training periods per week because some studies employed split programs. When split sessions were employed, different exercises were performed at different sessions. The training specificity principle implies that the training effect for a particular resistance exercise will derive almost entirely from the work done in those sessions in which that exercise is actually performed.

Sets per session. The number of sets per training session was coded for non-periodized programs. The number of sets per session varied from 1 to 6. When the sets per session varied over the course of a simple progressive program, the average was computed and rounded to the nearest whole number. The analyses of sets per session also included two specific contrasts that have been of interest in prior reviews: 1 set versus 3 sets (Galvao & Taaffe, 2004) and 1 set versus >1 set (Wolfe et al., 2004; Rhea et al., 2002). The number of sets was not coded for periodized programs. The systematic variation in the number of sets in periodized programs made it questionable whether any single value would be representative of the program.

Intensity. Intensity is the percentage of 1-RM lifted during each repetition in a training set. In practice, intensity is defined by the number of repetitions in a set. The assumption is that a person can complete one repetition for every 2 – 3% of 1-RM. Thus, defining the target intensity as 75% of 1-RM corresponds to the expectation that program participants will reach voluntary fatigue after completing 10 repetitions within a set of each exercise. If the target intensity was 80% of 1-RM, program participants would be expected to reach voluntary exhaustion after 8 repetitions. The translation of exercise repetitions into percentages is only approximate and some allowance must be made for fatigue that develops during a series of exercises. With this in mind, training intensity is usually defined as a range of repetitions, e.g, 8 – 10 repetitions per set.

Most studies covered in this review defined intensity in terms of the target range for the number of repetitions to voluntary exhaustion (e.g., 8–10 repetitions). Intensity was coded by taking the target range midpoint for repetitions. If the study employed a familiarization period, weighted average of the midpoints of the target ranges for the familiarization and training periods was computed. The weight was the number of weeks for each target range. The weighted averages were rounded to provide the final estimate of the number of repetitions per set. The rounded number of repetitions was converted to estimated percentages of 1-RM, subtracting 2.5% for each repetition. For example, if a program averaged 8 repetitions per set over the entire training period, the coded intensity was (1 – (8*.025)*100) = 80%. The estimated average intensity was coded into 5% ranges (e.g., 72.5% and 75%) with the upper bound used to label the range.

Periodization. Periodization is a training strategy that varies training objectives, training intensity, and training volume in planned cycles (Wathen, Baechle & Earle, 2000). The typical periodized program has three phases. The hypertrophy/endurance phase employs exercises with very low to moderate intensity (50% to 75% of 1-RM) with high to moderate volume (3-6 sets of 10 -20 repetitions. The basic strength phase couples high intensity exercise (80% - 90% of 1-RM) with moderate volume (3-5 sets of 4 – 8 repetitions). The strength/power phase combines high intensity (75% - 90% of 1-RM) with low volume (3 – 5 sets of 2 – 5 repetiations). Progression through these three phases constitutes a training cycle. The overall structure of a periodized program includes microcycles of 1 to 4 weeks, mesocycles of several weeks to several months, and macrocycles that typically last a year, but may be as long as 4 years. Most of the studies in this review lasted only a few months, so the typical program involved one cycle through all three periodization phases. A few studies included multiple microcycles by having very short phases (e.g., 1 day or 1 week).

Table 2 describes the distribution of the demographic variables and program characteristics in the overall data set.

Analysis Procedures

Every study included in this review used a pretest–posttest design. For this reason, methods described by Morris and DeShon (2002) were applied to compute appropriate ESs for repeated measures (simple ES_{RM}; see Appendix A). Subsequent steps were taken to solve two problems. First, in many studies, more than one strength test was administered to each sample. Consequently, there was more than one ES for each sample. The associated loss of ESs independence posed statistical problems (Gleser & Olkin, 1994). Second, exploratory analyses showed that ES_{RM} depended on which strength test was administered. Different programs could appear to produce stronger or weaker than average effects by choosing particular strength tests as outcome measures. These problems were addressed by identifying the 10 most frequently used strength tests. Analysis showed that this restriction largely eliminated the differences between tests, so treating each test as providing a separate estimate of a common training effect for each sample was reasonable. The ES_{RM} values for those tests were averaged to obtain a single *average* ES_{RM} for each sample (see Appendix B).[1]

---

[1] This review employed several different ES. An effect that is labeled “ES” refers to an ES computed without considering the repeated measures data structure, i.e., $ES = \frac{\overline{x}_{Post} - \overline{x}_{\Pr e}}{SD_{\Pr e}}$. An effect that is labeled “ES_{RM}” is the difference between post-training and pre-training scores with an adjustment for repeated measures. An effect that is labeled “average ES_{RM}” refers to a measure that is the average of the scores on 1 to 8 scores on frequently used tests (see Appendix B). An effect that is labeled “adjusted average ES_{RM}” refers to the average ES_{RM} adjusted for program length using an equation describing the association of program length with effect size based on all ES_{RM} values (see Equation 1). Finally, an effect that is labeled “population-adjusted average ES_{RM}” is the average ES_{RM} adjusted for program length using separate equations for trained and untrained samples (see Equations 2 and 3).

Table 2

General Structure of the Data Set

	No. Studies	No. Samples	No. ESs	ΣN[a]
Gender				
Men	117	185	371	2247
Women	37	53	162	706
Men and women	39	59	179	1540
Age group				
Younger	151	240	504	3704
Older	45	62	224	870
Training status				
Untrained	131	189	514	2935
Trained	56	100	179	1278
Sessions/week				
1	4	6	19	70
2	66	107	215	1980
3	98	171	453	2304
4	7	9	20	110
5	1	1	4	12

	No. Studies	No. Samples	No. ESs	ΣN[a]
Periodization				
No	108	186	514	2635
Yes	68	108	199	1878
Sets per session				
1	17	31	87	543
2	6	11	32	203
3	65	114	336	1533
4	11	12	21	145
5	10	11	22	145
6	2	4	10	39
Intensity				
<60%	2	4	15	57
60%	3	9	16	124
65%	2	3	9	33
70%	3	4	17	97
75%	45	69	186	986
80%	25	41	127	593
85%	26	38	63	500
Total	196	302	728	4574

Note. The total sample size for the moderator variables can be less than 4574 because some moderator variables could not be coded for some samples.

[a] ΣN is the cumulative sample size for each group.

The number of ESs that contributed to the average ES_{RM} varied from sample to sample. The number of common strength tests administered to different samples ranged from 0 to 8. The average ES_{RM} most often was based on 1 (k = 121) or 2 (k = 115) tests. The averages were based on 3 to 6 tests for 57 samples. Samples that had not performed any of the common tests (k = 4) were dropped from the analysis.

Meta-regression models evaluated potential moderator variables. A moderator was a demographic variable or a program element that might account for variation in ES_{RM}. The meta-regression analyses applied Hedges and Olkin's (1985) general methods. These methods included weighted analysis of variance (ANOVA) and weighted linear regression. The weight variable was the inverse of the estimated variance for ES_{RM}.

Moderators were evaluated in two steps. The first step was an overall test for a moderator effect. This test determined whether the adjusted ES_{RM} differed significantly across the moderator groups. The second step was taken only if there was a statistically significant moderator effect. Post hoc comparisons were conducted to determine which groups differed significantly. The moderator groups were rank ordered from largest to smallest average ES_{RM}. The group with the largest average ES_{RM} was adopted as the reference group. The first post hoc test compared the reference group with the group with the second largest average ES_{RM}. If these two groups differed significantly, the post hoc comparisons stopped at this point. If the two groups did not differ significantly, the group with the third-largest average was compared with the reference group. The comparisons continued down the ranked-ordered moderator groups until a significant difference was found. The comparisons stopped at that point, and all remaining groups were classified as differing significantly from the reference group.

Some post hoc comparison procedures required multiple significance tests. Performing multiple significance tests increased the probability that at least one comparison would be statistically significant by chance alone. A Bonferroni significance criterion was adopted to fix the analysis-wide probability of error at 5% or less. The post hoc procedures involved $j - 1$ comparisons for a moderator with j levels. The Bonferroni criterion for each moderator was $p_{critical} = .05/(j-1)$.

The post hoc comparisons identified equivalence sets. These sets consisted of the design option with the largest average effect plus the alternative options that were not significantly different from this reference value. The sets were equivalent in the sense that the alternative set options could not be confidently classified as less effective than the optimum design option based on the available evidence.

Large samples can produce significant results even for trivial differences (Rosenthal & Rosnow, 1984). To avoid mistaking sample size for explanatory power, the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) was adapted to provide an ES index for the moderator analyses. This index is the proportion of the greater-than chance variation in ES_{RM} accounted for by a moderator or set of moderators (Appendix C). Cohen's (1988) ES criteria were applied to characterize the TLI as indicating trivial, small, moderate, or large moderator effects.

Funnel plots were constructed to evaluate the potential effects of publication bias (Light & Pillemer, 1984). Egger, Smith, Schneider, and Minder's (1997) regression method was applied to obtain a formal statistical assessment of the hypothesis that publication bias. The file drawer problem was not examined because both the typical ES and the total number of studies were large. Under those circumstances, Rosenthal's (1979) file drawer criterion certainly would be satisfied.

All analyses were carried out with the computer program SPSS-PC (Version 17).

Results

Program Length Effect

Average ES_{RM} increased with program length. Preliminary analyses of this association compared linear, quadratic, logarithmic, power, and growth models as functional representations of the relationship of ES_{RM} with program length. The logarithmic model given as Equation 1 provided the best prediction of average ES_{RM}:

$$y = .412 + .545 * \ln(t)\,. \tag{1}$$

The correlation of average ES_{RM} with program length was modest (r = .21), but statistically significant (χ^2 = 96.88, 1 *df*, $p < .001$).

The linear form of the model could produce a mistaken impression. The intercept, .412, might be mistakenly interpreted as indicating that ES_{RM} was >0 at the beginning of training. This would be the usual interpretation of the intercept if the Equation 1 was a simple linear regression of ES_{RM} on weeks of training. The intercept did not have this interpretation because the predictor had been transformed. Solving the equation for ES_{RM} = 0, the estimated time to produce an effect of this size was 3.3 days. This estimate would correspond to 1 or 2 training sessions in a typical program.

Initial Moderator Analyses

The average ES_{RM} was the dependent variable for the initial moderator analyses. All of the potential moderator effects were statistically significant except for periodization. The χ^2 and TLI values in Table 3 showed that adjusting for program length generally reduced the strength of the moderator effects. The exceptions were the slightly larger χ^2 values for periodization and repetitions per set. Despite the general trend toward weaker effects, adjusting for program length only changed one conclusion regarding the presence of a moderator effect. Age significantly moderated the average ES_{RM} values, but was only marginally significant in the analysis of the adjusted average ES_{RM} values.

Table 3

Initial Tests for Moderator Effects

	Average ES_{RM}[a]				Adjusted average ES_{RM}[b]		
	χ^2	*df*[c]	Sig	TLI	χ^2	Sig	TLI
Age	39.39	1	.000	.016	3.73	.054	.000
Gender group	84.40	2	.000	.034	61.96	.000	.025
Men vs. women	49.31	1	.000	.031	18.42	.000	.010
Training status	445.73	1	.000	.225	395.47	.000	.215
Periodization	.83	1	.363	.000	1.02	.313	.000
Sessions per week	58.29	4	.000	.013	40.20	.000	.005
Sets per session	93.85	5	.000	.038	73.27	.000	.025
1 Set vs. 3 sets[d]	50.46	1	.000	.040	38.56	.000	.030
1 Set vs. multiple[d]	66.46	1	.000	.042	58.43	.000	.039
Repetitions per set	88.79	6	.000	.040	92.20	.000	.047

[a]The *average* ES_{RM} was based on each samples scores for a subset of the 10 most frequently used strength tests (see *Analysis Procedures*). [b]The *adjusted average* ES_{RM} was the average ES_{RM} corrected for program length based on Equation 1. [c]The degrees of freedom were the same for both analyses. [d]The single set comparisons were added to provide direct comparisons to prior reviews.

Training status was an especially strong moderator of ES_{RM} (see Table 3). The χ^2 for the training status moderator effect, $\chi^2 = 395.47$, was larger than the sum of the χ^2 values for all of the other moderators, $\Sigma\chi^2 = 387.79$. Periodization was noteworthy as the only moderator that was not significant in either analysis.

Isolating Demographic Moderator Effects

The initial bivariate moderator analyses were followed by more focused analyses designed to isolate the effects of specific demographic variables. These focused analyses were needed because age, gender, and training status were confounded in the data. For example, all of the samples with trained participants were younger, and nearly all of them were male. Given the strong association of training status with the adjusted average ES_{RM}, the confounding of training status with age and gender could bias the assessment of age and gender moderator effects. To control for this possible bias, the moderator effects of age, gender, and training status were re-evaluated with the other two demographic variables held constant. For example, the analysis was limited to young men when the effects of training status were evaluated. The restriction eliminated any possible effects of age and gender on the comparison.

Table 4 presents the results that were obtained when appropriate restrictions were introduced to isolate the effects of gender, training status, and age:

> ***Gender***. Gender did not affect the training response for younger untrained individuals ($\chi^2 = 1.06$, $p < .303$). Older men produced a significantly stronger training response than older women ($\chi^2 = 104.31$, 1 *df*, $p < .001$).

Table 4

Isolated Demographic Moderator Effects

Gender	***Men***	***k***	***Women***	***k***	χ^2	***Sig.***
Younger untrained	2.23	73	2.35	32	1.06	.303
Older	3.45	24	1.95	16	104.31	.000
Training status	***Untrained***	***k***	***Trained***	***k***	χ^2	***Sig.***
Young men	2.23	73	1.01	76	245.39	.000
Age	***Younger***	***k***	***Older***	***k***	χ^2	***Sig.***
Untrained men	2.25	73	3.45	24	98.04	.000
Untrained women	2.35	32	1.95	16	9.26	.003

Note. The *adjusted average* $\mathrm{ES}_{\mathrm{RM}}$ was the dependent variable for the analyses reported in this table.

> ***Training status***. The training response was much larger in samples of untrained individuals than trained individuals (χ^2 = 245.39, 1 *df*, *p* < .001).
>
> ***Age***. Older men produced significantly larger training effects than younger men (χ^2 = 98.04, 1 *df*, *p* < .001). Older women produced a significantly weaker training effect than younger women (χ^2 = 9.26, 1 *df*, *p* <.003).

The exceptionally large adjusted average $\mathrm{ES}_{\mathrm{RM}}$ for older men was central to both the gender and age moderator effects. If this group had been ignored, there would have been no substantial effects for either gender or age. If further study were to show that the available data overestimate the true response of older men, it would be appropriate to conclude that neither age nor gender is an important training response moderator

Interaction of Population with Program Design Facets

Training programs often are designed for a specific population. For this reason, it was important to ask whether program design facets had the same effect in different populations. This question was pursued with training status and gender as the possible program design effect moderators. Age was not considered in connection with this question because it was not clear whether analyses that included the evidence from samples of older men would yield meaningful. Separate analyses for men and women were not conducted because the sample sizes would have been too small to have confidence in the results.

Weighted ANOVAs assessed the combined effects of demographic variables (e.g., gender) and program elements (e.g., number of sets per session) with the remaining demographic variables held constant. For example, Table 5 presents the results from analyses that were restricted to young men.

Table 5

Younger Untrained Males versus Younger Trained Males

	Experience		Design		ExD Interaction		ExD	Residual	
	χ^2	*df*	χ^2	*df*	χ^2	*df*	TLI	χ^2	*df*
Periodization	176.52	1	4.20	1	11.24	1	.028		142
Sessions	143.04	1	14.59	4	12.71	1	.031		136
Sets	69.96	1	6.29	5	16.69	2	.037		68
1 vs. 3 sets	149.44	1	1.11	1	1.78	1	.000		46
1 vs. multiple	164.70	1	1.58	1	8.69	1	.020		73
Repetitions	183.21	1	20.33	6	1.58	2	.000		62

Note. The dependent variable was the adjusted average ES_{RM}. The TLI values are partial TLI s based on the χ^2 values for the interaction and the residual.

The interactions of demographic characteristics with design facets were the primary concern in the analyses reported in Tables 5 and 6. A significant interaction indicated that the program facet impact depended on which population was considered. A significant interactions would be reason to consider providing population-specific training recommendations. Thus, the TLI for each interaction has been reported to ensure that the differences involved were large enough to be important.

Training status. For young men, the interaction of training status with program design was statistically significant for every element except the number of sets per session (see Table 5). The specific contrast of 1 set per session with 3 sets per session also failed to reach statistical significance ($\chi^2 = 1.78$, 1 *df*, $p > .182$), but the overall effect of sets per session was significant ($\chi^2 = 16.69$, 2 *df*, $p < .001$) as was the contrast of 1 set with >1 set per session $\chi^2 = 8.69$, 1 *df*, $p < .003$). In each case, program design affected training outcomes for untrained individuals more than it affected the training outcomes for trained individuals.

Table 6

Gender Holding Age and Experience Constant

	Gender		Design		GxD Interaction		GxD	Residual	
	χ^2	*df*	χ^2	*df*	χ^2	*df*	TLI	χ^2	*df*
Periodized	.10	1	9.63	1	.00	1	.000	454.47	98
Sessions	4.55	1	35.86	3	.37	1	.000	419.34	94
Sets	.02	1	11.58	5	1.39	1	.000	371.16	69
1 vs. 3 sets	.31	1	.77	1	1.20	1	.000	332.41	56
1 vs. multiple	.20	1	.00	1	1.58	1	.000	382.50	74
Repetitions	.22	1	14.66	5	3.88	4	.000	304.18	58

Note. Analyses compared young untrained men and young untrained women. The dependent variable was the adjusted average ES_{RM}.

Gender. The gender analysis compared young untrained men and women. No gender interaction was statistically significant (see Table 6). In every case, the variation explained by the interaction was not even as large as would have been expected by chance (i.e., TLI = .000). It was also worth noting that gender produced a statistically significant difference in only one of six analyses (sessions per week).

Population-Specific Adjustments for Program Length

Population-specific adjustment equations. Population-specific program length adjustments were explored because the weighted average adjusted ES_{RM} of trained individuals was less than half that of untrained individuals (trained, adjusted average ES_{RM} =.97; untrained, adjusted average ES_{RM} = 2.24). The two populations underwent training programs of approximately equal length (trained, 10.84 weeks; untrained, 10.28 weeks; $t_{1,222} = 1.10$, $p = .272$), so the difference in the average training effect sizes implied different growth rates for training effects.

The relationship of program length, expressed in weeks, to ES_{RM} depended on training status. The regression lines were not parallel in an analysis of covariance (χ^2 = 13.51, 1 *df*, $p < .001$). The population-specific regression equation for ***untrained individuals*** was:

$$ES_{RM} = .111 + .781 * \ln(length)\,. \qquad (2).$$

The relationship was highly significant (χ^2 = 57.48, 1 *df*, $p < .001$, $r = .29$) in this population. The corresponding regression equation for ***trained individuals*** was:

$$ES_{RM} = .485 + .189 * \ln(length)\,. \qquad (3).$$

The relationship was not statistically significant (χ^2 = 1.21, 1 *df*, $p >.271$, $r = .07$) for trained individuals. Solving the equations for ES_{RM} = .00 indicated that positive training effects would be predicted after 6 days for untrained individuals and after one day for trained individuals. The former value implies that 3 or 4 training sessions are needed before the ES_{RM} exceeds 0; the latter value implies gains from the first training session onward.

The smaller coefficient for ln (length) in Equation 3 for trained individuals indicated a slower rate of improvement for that population. The association was not statistically significant, but it was consistent with everyday observations that even trained populations show strength gains over time. For this reason, population-specific adjusted ES_{RM} variables were computed for both trained and untrained samples. In the following results, average ES_{RM} estimates that have been adjusted using Equation 2 or Equation 3 are referred to as *population-adjusted average* ES_{RM} to distinguish them from the earlier adjusted average ES_{RM} estimates based on Equation 1.

Population-specific moderator effects for untrained individuals. The earlier assessment of training status as a moderator was limited to young men. Samples were included regardless of gender composition because the population-specific moderator analyses had shown that gender did not moderate the training response for the adjusted average ES_{RM}.

Table 7

Design Facet Moderator Effects for Young Untrained Individuals

Moderator	Level	Population-adjusted avg[a]	k[b]	Equivalence set[c]
Periodized	No	2.34	100	
	Yes	2.01	24	
		$\chi^2 = 9.85$, *df* =1, *p* <.001, TLI = .009		{No}
Sessions	1	0.99	1	
	2	1.96	26	
	3	2.42	93	
	4	1.67	1	
	5	2.29	1	
		$\chi^2 = 36.51$, 4 *df*, *p* < .001, TLI = .023		{3}
Sets	1	1.96	18	
	2	2.39	6	
	3	2.49	56	
	4	2.38	10	
	5	1.76	3	
	6	2.38	4	
		$\chi^2 = 24.87$, 5 *df*, *p* < .001, TLI = .000		{3, 2, 6, 4}
Intensity	< 60%	2.49		
	65%	1.58		
	70%	2.57		
	75%	2.45		
	80%	2.01		
	85%	2.79		
		$\chi^2 = 39.31$, 5 *df*, *p* < .001, TLI = .071		{85%, <60%)

[a] See text for the definition of the population-adjusted average ES_{RM}. [b] "*k*" is the number of samples that provided averages for analysis. [c] The equivalence sets include all design options that were not significantly different from the option with the highest population-adjusted average. The design options have been listed from largest to smallest population-adjusted average ES_{RM} in the set.

All four program design facets proved to be statistically significant moderators of the population-adjusted average ES_{RM} training effect for untrained program participants (see Table 7).

Periodization. Periodized programs were less effective than simple progressive programs, but the difference—while statistically significant—was too small to be important (TLI = .009). Non-periodized programs were singled out as a best practice, but the small TLI made this a dubious designation.

Sessions per week. The post hoc comparison was limited to contrasting 2 sessions per week with 3 sessions per week. No comparisons were made for 1, 4, or 5 sessions per week because each of these options was represented by a single sample. Therefore, while the post hoc comparisons identified 3 sessions per week as a best practice for this population, this designation was advanced with the caution that this characterization should not be given too much credence until there is more evidence regarding other options—especially 4 and 5 sessions per week.

Sets per session. The equivalence set included four design options. Three sets per session produced the largest population-adjusted average ES_{RM}, but that mean effect of 3 sets per session was not significantly greater than mean effects seen with 2, 4, or 6 sets per session. Also, 1 set per session was significantly less effective than other choices, whether the comparison was to all other options

($\chi^2 = 22.00$, 1 *df*, $p < .001$) or to the specific choice of 3 sets per session

($\chi^2 = 19.24$, 1 *df*, $p < .001$).

Intensity. An 85% target intensity produced the largest population-adjusted average effect. The post hoc comparisons added 75% or <60% to the equivalence set. The differences were statistically significant even when the analysis was limited to the three intensities with >3 population-adjusted average ES_{RM} values ($\chi^2 = 32.37$, 2 *df*, $p < .001$).

Moderator effects for trained individuals. The population-adjusted average ES_{RM} based on Equation 3 was the dependent variable for the analyses reported in Table 8.

Periodization. Periodized programs produced slightly larger effects than non-periodized programs, but the difference was not statistically significant.

Sessions per week. The initial moderator test was not statistically significant, so the equivalence set included all four options.

Sets per session. Four sets per session produced the largest training effect. Post hoc comparisons indicated that the effect of 4 sets per session was not significantly greater than the effect of 5 sets per session. The difference between 1 and 3 sets per session approached statistical significance ($\chi^2 = 2.65$, 1 *df*, $p < .104$). The difference between 1 set and >1 set was statistically significant ($\chi^2 = 7.07$, 1 *df*, $p < .001$).

Intensity. A target intensity of 60% produced the largest effect, but it was not significantly greater than the effect for 80% or 85%. Note that if the effect for 60% is ignored, the trends for intensity followed the general rule of thumb that “more is better.” The population-adjusted average ES_{RM} increased from 75% to 85%, but those three options did not differ significantly ($p > .094$). Thus, the evidence would support an equivalence set consisting of three options, even if the 60% programs were excluded from consideration.

Table 8

Design Facet Moderator Effects for Young Trained Individuals

Moderator	Level	Population-adjusted avg.[a]	k[b]	Equivalence set[c]
Periodized	No	.84	33	
	Yes	1.01	56	{Yes, No}
		$\chi^2 = 2.82$, 1 *df*, $p > .093$, TLI = .012		
Sessions	1	.74	5	
	2	.92	55	
	3	1.00	25	
	4	.93	5	
		$\chi^2 = 1.41$, 3 *df*, $p > .703$, TLI = .000		{3.4, 2, 1}
Sets	1	.46	7	
	3	.78	16	
	4	1.73	2	
	5	1.03	7	
		$\chi^2 = 15.72$, 3 *df*, $p < .002$; TLI = .183		{4, 5}
Intensity	60%	1.41	6	
	75%	.58	13	
	80%	.78	3	
	85%	1.04	8	
		$\chi^2 = 15.95$, 3 *df*, $p < .002$, TLI = .136		{60%,85%}

[a] See text for the definition of the population-adjusted average $ES_{RM.}$ [b] "k" is the number of samples that provided averages for analysis. [c] The equivalence sets include all design options that were not significantly different from the option with the highest population-adjusted . The design options are listed from largest to smallest population-adjusted ES_{RM} in the set.

Publication Bias

Separate funnel plots (Light & Pillemer, 1984) were constructed for the adjusted average ES_{RM} for young untrained, young trained, and older program participants. The plots appeared to be truncated on the left hand side because even small studies always produced positive effects. Symmetry would have required some small to moderate negative effects. The regression method developed by Egger et al. (1997) provided statistical confirmation of the impression derived from the funnel plots. The intercept of the regression was significantly ($p < .001$) greater than zero in each population.

Discussion

This review did not identify best practices. Program design facets were statistically significant moderators of ES_{RM}, but post hoc analyses seldom singled out one option as significantly better than all others. Only one case involving more than two options produced a best practice. Four sets per session was the best option for this design facet for trained individuals. Rhea et al. (2003) reached the same conclusion, but it should be noted that the estimated effect of 4 sets per session has been based on very little evidence. In the present review, the estimate was based on evidence from just two samples. Even one or two additional studies that produced smaller effects for this option could alter the conclusion that 4 sets per session is a best practice.

It was also true that non-periodized training programs were more effective than periodized programs for untrained individuals. Non-periodized training programs technically would be a best practice, but the TLI was too small to make this a strong recommendation. Such guidance would not really qualify as designating a best practice in any case. The design of a non-periodized program requires decisions about the number of sessions, sets per session, and repetitions per set. Best practices could not be identified for those program design facets, so knowing that non-periodized programs were better than periodized programs would not lead to the specification of any optimal design specifics.

Failing to identify best practices is not unique to this review. Other reviews show the same general pattern of results. Program design facets often are statistically significant moderators of the training response, but post hoc analyses fail to identify any single option as significantly better than all other options (see Appendix E). Given this general trend, the current findings could not be dismissed as resulting from the inclusion criteria or analysis procedures that have been employed in the current review.

The statistical methods adopted in this review should have sharpened the contrasts between design options. First, repeated measures analyses are expected to produce larger effect sizes. This effect was present in the current analyses as indicated by comparing the average ES_{RM} of 1.76 with the average simple ES of 1.05. The difference arose because the estimated sampling variability was smaller for repeated measures analyses. The smaller sampling variability also would amplify differences between the average ES_{RM} values for different design options in post hoc analyses, so the current procedure should have increased the likelihood of finding a best practice. Second, if the results obtained with different strength tests all are estimates of the same training effect, the use of average ES_{RM} measures increased the precision of the effect size estimates. Increasing precision is the same as increasing measurement reliability (American Psychological Association, 1985), thereby reducing the effects of attenuation due to measurement error (Nunnally & Bernstein, 1994). Once again, this should increase the likelihood that the post hoc comparisons would be statistically significant. Finally, adjusting for program length differences removed a source of variance that otherwise could have obscured differences between treatment options.

The failure to identify best practices does not mean that such practices do not exist. Every analysis produced one option that had a larger average ES_{RM} than all other options for that facet. The problem was that the differences between the most promising option and other choices were not large enough to be statistically significant. Although the comparisons have not been reported in detail here, many post hoc comparisons produced very small χ^2 values despite moderately large sample sizes. The implication is that the available evidence would have to be multiplied many times to make the contrasts between the design options statistically significant. If the required data were available, the conclusion still might be that the differences were too small to be important. That argument would be particularly powerful when coupled with the knowledge that training effects increase with program length. In the final analysis, the difference between design options might amount to choosing between one program that would produce a given effect in 10 weeks and another program that would require 12 weeks to produce the same results. It is debatable whether the extensive additional research that would be needed to clearly define best options would really have much impact on program design choices.

A low probability of identifying best practices at any time in the near future does not mean that resistance training research has no value. Research has singled out some design options as less effective than others. If the typical equivalence set included more than one option, it is also true that it seldom contained all possible options. For example, the overall evidence justified ruling out the use of single set programs unless in the absence of design constraints that make it impossible to incorporate more than one set per session. This point has been debated in the past, but the aggregate body of evidence summarized here and in other recent reviews has reached the point that there is little doubt that multiple-set programs are superior to single-set programs. This result suggests a guideline for future studies. It may be more productive to undertake studies designed to rule out some options than to focus attention on identifying the best option. Here again, though, the ultimate difference may be that the less effective options simply take longer to reach program targets.

Research to date has identified only one strong influence on the training response. Untrained individuals have produced much greater training effects than trained individuals (Rhea et al., 2003; Payne et al., 1997; Rhea et al., 2002; Wolfe et al., 2004). Gender has had little impact on the training response—a finding that corroborated earlier reviews that used different inclusion criteria and analytic methods (Payne et al., 1997; Peterson et al., 2004; Rhea et al., 2003; Rhea & Alderman, 2004; Wolfe et al., 2004). Whether age affects the training response is uncertain. The age effects for men and women were in the opposite direction, but the possibility cannot be ruled out that this result was due to a few atypical male samples.

The smaller average response of trained individuals deserves further comment. Anecdotal evidence and the flatter slope of the program length-ES_{RM} equation for trained individuals both suggest a slower rate of improvement for trained individuals. The smaller average response of trained individuals has the same implication because the average program length was virtually identical for trained and untrained samples. This rate difference has implications for evaluating training programs that have been ignored to date. Training programs’ effects accumulate over time, so research that compares different programs must continue long enough to show the differences in the cumulative training effects. Longer studies are needed to accurately evaluate different program designs in trained populations.

This review has provided additional perspective on some ongoing resistance training controversies. The value of single-set programs has been debated with some reviewers favoring the conclusion that single-set programs are as good as multiple-set programs (Carpinelli & Otto, 1998), while others favor the conclusion that multiple-set programs are more effective (Galvao & Taaffe, 2004; Rhea et al., 2002; Wolfe et al., 2004). This summary of the evidence came down in favor of multiple-set programs.

This review has also provided additional perspective on periodized training programs. Periodized programs were slightly less effective than non-periodized programs for untrained individuals. The two approaches produced comparable results in trained individuals. These results appear to conflict with Rhea and Alderman's (2004) conclusion that periodized programs are effective regardless of training status, age, or gender. The key to the apparent conflict may be that there is a difference between knowing that a program is effective and knowing that it is more effective than some alternative program. Rhea and Alderman (2004) analyzed their data in two phases. First, they showed that periodized programs were more effective than non-periodized programs when all of their effect sizes were analyzed together. Subsequently, Rhea and Alderman (2004) showed that periodized programs produced statistically significant gains for men and women, young and old, and trained and untrained individuals. The results of the second phase of the analyses only lead to the conclusion that periodized programs are superior to non-periodized programs for all types of people if the difference observed in the initial analysis applies equally to all subgroups. The present analyses suggested that this generalization cannot be taken for granted. Indeed, the findings in this review raised doubts about the overall superiority of periodized programs. Further investigation of this topic could be fruitful, but it is worth noting that identifying a difference between periodized programs and non-periodized programs will not simplify the problem of defining best practices. Appropriate choices would still be needed for the number of sessions per week, sets per session, and repetitions per set will have to be made no matter which approach is ultimately identified as the better option. Evidence favoring periodization would amplify these requirements by making it necessary to specify a choice for each facet in each microcycle, specify the duration of each microcycle, and so forth.

Limitations of this review must be considered when evaluating the findings. No attempt was made to conduct ancestry searches or to systematically search for unpublished studies, so it is likely that the total volume of evidence could have been increased. However, unless the omitted literature is very large, additional evidence would be very unlikely to change this review's primary conclusions. Furthermore, some program design options have been studied so infrequently (e.g., 4 sessions per week for untrained individuals) that it is unlikely that any search would yield enough evidence to reach strong conclusions about the utility of those options. A third important point is that this review relied on estimates of the correlation of pre-training test scores with post-training test scores. Estimates had to be used because training studies rarely report this statistic or other statistics from which it could be derived (e.g., a sample-specific correlated *t* test). Another point to consider is that the analyses relied on a fixed-effects statistical model. Large residual χ^2 values indicated that adopting a random-effects model would have been a reasonable course of action. However, shifting to a random-effects model would only have accentuated the central finding of the study. The fixed-effects model overestimates the estimated effect size precision (National Research Council, 1992). Shifting from a fixed-effects model to a random-effects model would have increased the variance estimate for individual effect sizes. The increased variance would have reduced the χ^2 values in the analyses. Any difference between design options that was not statistically significant in the fixed-effects analysis would be even less likely to be significant in the random-effects model. The use of a random-effects model would have been expected to increase the size of the equivalence sets that were the central products of this review. A fifth issue to consider is that no attempt was made to correct for publication bias. Corrections would have resulted in a somewhat smaller overall effect size estimate, but the tests for publication bias may have been misleading because bias is not the only process that could have generated the same pattern of findings (Tang & Liu, 2000). Finally, the interactions between program design facets were not investigated. A meaningful evaluation of those interactions would require data from studies that represented a wide range of the possible design facet combinations. The fact that some individual design options have only been studied in a single sample made it clear that research to date has not provided the empirical evidence needed to fully evaluate interactions.

What guidelines are appropriate for designing resistance training programs? The present failure to identify best practices left this question unanswered, but the findings provide a frame of reference for an answer. The key observations are that it will usually be the case that several design options will produce similar effects and that training status is important. It follows that guidelines should present a range of alternatives for the number of sessions per week, the number of sets per session, and the intensity of repetitions within sets. The choices among alternatives should be tailored to the training status of the program participants. Guidelines that meet these criteria and embody informed professional judgments based on the empirical evidence are already available (American College of Sports Medicine, 2009; Kraemer et al., 2002). The best recommendation for program design at this time is to follow those guidelines. It is unlikely that better guidelines will be available in the near future given the difficulty of providing convincing evidence that options within equivalence sets produce demonstrably different training outcomes.

Appendix A

Computing Effect Sizes for Repeated Measures

Meta-analysis provides estimates of the average ES and the variation of individual ES estimates about that average. The homogeneity tests for variation about the average are especially important in the present context. If the ESs for different training programs display greater-than chance variation, it is reasonable to search for moderator variables that can explain the observed heterogeneity. In the present review, program design facets and demographic variables were of interest as potential moderator variables.

Studies must be assigned appropriate weights to compute the average ES and test for variation about the average. The weights are based on the precision of the individual ES estimates. All studies reviewed here employed pretest–posttest research designs. In such cases, the correlation of pretest scores with posttest scores affects the sampling variance that is the ES estimate index of precision. Therefore, the pretest–posttest correlation must be known to derive sampling variance estimates that are suitable for determining ES weights. The correlation must be known whether the analyses employ standardized mean change scores or difference scores (Morris, 2000). For change scores, the proper estimate of sample variance is:

$$\sigma_{Diff}{}^{2} = \sigma_{1}{}^{2} + \sigma_{2}{}^{2} - 2r\sigma_{1}\sigma_{2} \tag{A1}$$

In this equation, the subscripted “Diff” indicates that the variable of interest is a difference score. The pretest–posttest correlation, *r*, is expected to be positive and moderate to large. As a consequence, the last term of Equation A1 will be moderate to large relative to the first two terms. It follows that simply pooling the pretest and posttest variances, as would be the case if the pretest–posttest correlation was ignored, will result in overestimation of the true sampling variance. If the variance is overestimated, the *z* scores associated with the deviation of specific ES values from the average ES will be smaller than they would if the correct variance were used. The overall test for homogeneity of ESs, Cochran’s *Q*, is the sum of the squared *z* scores. Thus, overestimating sampling variance will lead to underestimating *Q*. This bias in the *Q* test values could lead to the erroneous conclusion that a given moderator is unimportant. The tests for moderators were central to this review, so accurate variance estimates were essential.

The correct variance estimates could be estimated easily if studies routinely reported the pre-training/post-training correlations for test scores. Unfortunately, this information is seldom reported. The required information could be extracted from the *t* tests or *F* tests for the time effect if either statistic was reported separately for each condition in the study. Once again, resistance training studies seldom provide this information.

Similar problems have been encountered in other meta-analytic contexts. An analogous problem arises when research syntheses must adjust for measurement error (Hunter & Schmidt, 1990). Because reliability estimates are reported only infrequently for the primary studies in an analysis, reliability estimates are obtained by averaging those that have been reported (e.g., Safrit, Hooper, Ehlert, Costa, & Patterson, 1988). The same approach can be applied to the problem of estimating the pretest–posttest correlations for resistance training studies.

Table A1 presents average pretest–posttest correlations for strength tests administered in several resistance training studies. The averages are derived from the information in studies by Hagerman (2001), Karabulut (2008), Konstanty (1990), Lucier (1999), Omizo (1992), Van Oosbree (1993), and Womer (2003). These studies reported the correlation, the *t* test or *F* test for the time effect, or the raw data needed to compute these statistics. The averages also included the results of analyzing data from several studies of Navy personnel (Marcinik, Hodgdon, Englund, & O'Brien, 1987; Marcinik, Hodgdon, & Vickers, 1985; Marcinik, 1986; Marcinik, unpublished manuscript). Table A1 reports the resulting weighted averages for 6 strength tests that were represented by at least three estimates of the test-retest correlation. The cumulative sample sizes for those tests ranged from 290 to 459 observations.

Table A1

Test-by-Test Estimates of the Pooled Pre-Post Correlation

Test	k	∑N	Mean r_{12}
Bench	10	497	.909
Shoulder press	4	344	.802
Lat pull-down	3	290	.826
Biceps curl	3	333	.770
Leg press	6	459	.817
Knee extension	4	112	.723
Total		2120	.834

A meta-analysis of the differences in the correlation coefficients for the various strength tests was conducted using the Fisher *r*-to-*Z* transformation. The analysis showed that the reliability estimates for the six tests in Table A-1 differed significantly (χ^2 = 71.89, 5 *df*, $p < .001$). However, most of the variation was attributable to the difference between the average correlation for bench press and the corresponding statistics for the remaining five tests. When the bench press was eliminated from the analysis, the remaining differences were not statistically significant (χ^2 = 8.04, 4 *df*, $p > .090$). The average test-retest correlation in this subset of the strength tests was $r = .800$, a value that was representative of all five tests. Based on these findings, repeated-measures effect size computations proceeded with the test-retest correlation for the bench press set at $r = .90$, and the test-retest correlation estimate for other strength measures set at $r = .80$.

After developing estimates of the pretest–posttest correlations, the analysis followed guidelines provided by Morris and DeShon (2002). First, the variance for individual observations was computed by applying Equation A-1 above. Second, the standard deviation of the differences (SD_{diff}) was computed by taking the square root of the variance. This standard deviation was used to compute the initial ES_{RM} (Equation

A2). A separate ES was computed for each record in the data file. A record consisted of the results for a single strength test administered to a particular subject sample.

The use of an average pretest–posttest correlation in many cases will obviously be inaccurate. However, these correlations clearly have been positive and substantial when estimates have been available. Ignoring this strong trend would lead to very conservative tests for moderator effects. The uncertainty introduced by the use of average values was preferable to having results that certainly were too conservative.

The estimated pretest–posttest correlation values were combined with the sample standard deviations to compute the variance of the difference scores as shown in Equation A1. The standard deviation of the difference (SD_{diff}) was the square root of this variance. The ES for repeated measures was:

$$ES_{RM} = (Mean_{post} - Mean_{pre}) / SD_{diff} \tag{A2}$$

Weighting ES_{RM} estimates. Individual ES estimates must be weighted to obtain the most precise aggregated $\mathrm{ES_{RM}}$ estimate and to test for heterogeneity in the individual estimates. The appropriate weights are the inverse of the variance. The variance of an individual $\mathrm{ES_{RM}}$ estimate can be computed by applying the equation for the single-group pretest–posttest change score variance formula in Table 2 of Morris and DeShon (2002, p. 117).

$$Variance = \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1 + n\delta_{RM}{}^2\right) - \left(\frac{\delta_{RM}{}^2}{c^2}\right) \tag{A3}$$

In this equation, n is sample size and δ_{RM} is the population value for $\mathrm{ES_{RM}}$. The equation includes a bias correction, c, to obtain accurate variance estimates. This correction factor was obtained by applying the approximation developed by Hedges (1982) and given as Equation 23 in Morris and DeShon (2002, p. 117).

$$c = 1 - \frac{3}{(4 * df) - 1} \tag{A4}$$

The variance computations required one additional input, δ_{RM}. Ideally, this parameter would be set equal to the unknown population ES. An estimate of this population parameter, d_{RM}, was used because the population value can only be estimated once after the variance is already known. Given this circularity, the recommended solution is to compute the unweighted ES and use that value for computing the variance for $\mathrm{ES_{RM}}$ (Hedges, 1982; Morris & DeShon, 2002). The present analyses employed this approach.

The variance of $\mathrm{ES_{RM}}$ was computed by applying Equation A3 after estimating δ_{RM}. The derivation of that equation can be found in the appendix to Morris and DeShon (2002) or in Gibbons, Hedeker, and Davis (1993). The accuracy of the syntax used to implement the equation in the present analyses was confirmed by repeating the small meta-analysis given in Table 3 of Morris and DeShon (2002).

Appendix B

Selection of Strength Tests for Analysis

For every sample, variants of the average ES_{RM} were the dependent variables in the analyses reported in this paper's main body. This appendix describes the analyses that were undertaken to determine whether averaging was appropriate.

Table B1

Average ES_{RM} for Different Strength Tests

	No. Effects	Cumulative Scores	Mean	SD	SE	Min	Max
Bench Press	161	2336	1.53	1.15	0.09	-0.28	6.40
Leg Press	91	1248	1.98	1.48	0.16	0.18	7.07
Squat	84	1063	1.56	0.95	0.10	0.07	4.22
Knee/Leg Extension	81	1038	2.16	1.27	0.14	0.23	6.09
Biceps Curl	63	1425	1.61	1.05	0.13	0.27	5.20
Leg Curl	44	630	1.71	0.93	0.14	0.09	3.77
Chest Press	38	588	2.21	1.95	0.32	0.27	9.70
Triceps/Elbow Ext	32	434	2.24	1.35	0.24	0.66	5.94
Lat Pull-down	28	392	2.27	1.48	0.28	0.76	7.00
Mil/Shoulder Press	20	252	1.80	1.41	0.32	0.32	5.45
Miscellaneous Rowing	14	241	2.41	1.50	0.40	0.83	6.46
Miscellaneous Core	13	168	1.29	0.52	0.14	0.43	2.01
Power Clean	9	74	0.23	0.43	0.14	-0.12	1.33
Lateral Raise	7	103	1.40	0.33	0.12	0.76	1.67
Calf Raise	6	98	2.72	1.37	0.56	1.14	4.89
Hip Abduction	6	86	2.54	0.87	0.36	1.41	3.51
Hip Extension	6	85	3.11	1.32	0.54	1.75	5.14
Hip Adduction	5	75	2.42	1.02	0.46	1.02	3.67
Hip Flexion	4	63	3.59	0.66	0.33	2.80	4.33
Chest Fly	1	8	2.20	N/A	N/A	2.20	2.20
Total	713	10407	1.83	1.29	0.05	-0.28	9.70

Initial assessment of test effects on ES_{RM}. The analyses began with an examination of the unweighted ES_{RM} values for 18 specific strength tests and two categories of tests that were comparable in general intent, but differed in the specifics of the exercise (see Table B1). The goal of this analysis phase was to determine whether a single average value could be used in the procedures described in Appendix A to compute the variance estimates for each individual test regardless of which test was considered. If the average ES_{RM} for individual tests varied significantly across tests, the use of a test-specific average would be more reasonable.

The initial one-way ANOVA indicated that a single value was not suitable. The average ES_{RM} for different tests varied from a minimum of 0.23 for power clean to a maximum 3.59 for hip extension. Test differences were statistically significant ($F_{19,693} = 3.56$, $p < .001$) and accounted for 8.1% of the total variation in ES_{RM} (see Table B2).

Table B2

Impact of Strength Test on Effect Size Estimates

All tests	*SS*	*df*	MS	*F*	Sig	ε^2
Between	116.69	19	6.14	3.96	.000	.081
Within	1075.31	693	1.55			
Common tests						
Between	52.01	9	5.78	3.60	.000	.037
Within	1013.32	632	1.60			

Note. Common tests were those represented 20 or more effect sizes in the data. Only four samples that produced a total of 6 ESs were lost from the analysis because they had not performed any of the common tests.

The average values for individual tests gave reason to believe that a subset of the tests might have unduly inflated the overall differences. Eight questionable tests were represented by <10 ES_{RM} values (see Table B1). Seven of those eight tests produced extreme averages, including the smallest average ES_{RM} and the six largest average ES_{RM} values in the table.

Another comparison shed further light on the impact of the number of ES_{RM} estimates available to describe a given strength test. The average ES_{RM} for tests with ≥20 ES_{RM} values ranged from 1.53 to 2.27. Only one of the eight tests represented by <20 ES_{RM} values produced an average ES_{RM} within this range. The other 9 averages were either less than the lower bound of this range or greater than the upper bound. The test with an average ES_{RM} within the range, chest fly, was represented by a single sample comprised of eight subjects. The average ES_{RM} value for the two miscellaneous categories also fell outside the range of averages for the frequently used tests.

Focus on frequently used tests. The second analysis phase examined the feasibility of limiting the coverage in this review to a subset of strength tests. The rationale for this step was that the extreme values observed for tests that had been administered to <20 samples might be misleading because there was too little data to obtain accurate estimates for those tests. Retaining the extreme averages could exaggerate the true magnitude of test differences.

The second analysis examined differences between the 10 most common strength tests. The test-to-test variation in test scores was significant even when the analysis was restricted to the 10 most common tests (see Table B2). The proportion of variance explained by the test differences was reduced (3.7% vs. 8.1%), but so were the degrees of freedom associated with this variance (9 vs. 19). The percentage of variance per degree of freedom remained approximately constant.

The ANOVA results provided the basis for assessing the variation in ES_{RM}. The assessment used the variance explained and the number of degrees of freedom to compute an effect size for the test-to-test differences. The differences failed to reach the minimum value that Cohen (1988) would classify as a small effect in either the analysis based on 20 tests or the analysis based on 10 tests. However, the ε^2 for the analysis based on 10 frequently used tests was clearly below the criterion value for a small effect while the ε^2 for the analysis based on all 20 tests was quite close to the lower bound for a small effect. This difference focused subsequent attention on the 10 common tests.

At this point in the analyses, it was judged appropriate to use test-specific averages when computing ES_{RM} variance. Thus, preliminary analyses were conducted with weights based on variances computed with a test-specific d_{RM} for each $ES_{RM.}$ The results of those preliminary analyses were the point of departure for a third assessment of the need for test-specific averages.

Table B3

Strength Test Differences in Specific Populations

	SS	*df*	MS	*F*	Sig.	ε^2
Young untrained women						
Between	51.15	9	5.68	3.19	.003	.222
Within	106.98	60	1.78			
Young untrained men						
Between	20.14	9	2.24	1.73	.088	.042
Within	181.53	140	1.30			
Young trained men						
Between	3.24	8	.40	.46	.882	.000
Within	108.72	123	.88			
Older						
Between	8.26	9	.92	.50	.873	.000
Within	314.21	171	1.84			
Young untrained women w/o extremes						
Between	16.85	9	1.87	1.82	.084	.089
Within	58.64	57	1.03			

Training population effects. A third analysis was undertaken after preliminary analyses using weights based on test-specific averages showed that demographic variables were strongly related to ES_{RM}. Given those strong effects, any demographic variable confounding with strength test usage could distort the effects of strength tests on ES_{RM}. To determine whether strength tests produced different average ES_{RM} values when demographic variables were held constant, the strength test comparisons were repeated in four homogenous subgroups: untrained young men, untrained young women, trained young men, and older populations. The analysis was restricted to the 10 most frequently used tests to ensure that the test results were comparable across populations (see Table B3).

Table B4

Reliabilities Based on Intraclass Correlations

No. tests	No. samples	Estimated ICC[a]
1	117	.635[b]
2	116	.777
3	24	.385
4	15	.731
5	9	.632
6	9	.461

Note. The total number of samples in these analyses was only 290 because intraclass correlations were not computed for samples with 7 test scores (k = 4) or eight test scores (k = 4). The sample sizes for those analyses were judged too small to obtain reasonable estimates of the typical difference between samples.

[a] Intraclass correlation coefficient [b]The estimated single test ICC was derived by applying the Spearman-Brown prophecy formula (Nunnally & Bernstein, 1994, pp. 262-264) to the ICC for samples with two tests.

Strength test was not a significant predictor of ES_{RM} in three of four demographic groups. Using the method of adding *p*s (Rosenthal, 1978), the pooled probability for the four groups was $p = .484$. The data for young, untrained women data then were reanalyzed with three extreme effects sizes ($ES_{RM} > 5.80$) dropped from the analysis. The differences for this group no longer were statistically significant ($p = .084$) and the pooled probability increased ($p = .575$).

Further consideration of the results from the subgroup analyses focused on the variance explained by strength test differences. The variance explained failed to meet the minimum criterion for even a small effect size for the untrained young men, trained young men, and older populations. The variance explained for the untrained young women exceeded the minimum value for a small effect in the initial analyses, but fell just below this minimum when the three extreme samples were excluded.

The preceding analyses indicated that there was no reason to adopt different d_{RM} values for different strength tests provided attention was limited to the 10 most frequently used tests. An average ES_{RM} based on the 10 frequently used tests was a reasonable index of the impact of a given training program.

Reliability of the average ES_{RM}. If the results for different strength tests provide estimates of a common ES_{RM}, the average ES_{RM} should be more reliable than a single score. Intraclass correlation coefficients were computed to determine the reliability of the aggregated score (see Table B4). The typical value indicated low to moderate reliability as reflected in the weighted average of $r_{xx} = .670$.

Impact of test selection. The analyses reported in this paper were restricted to the average ES_{RM} for whichever of the 10 strength tests represented by ≥ 20 ES_{RM} estimates had been administered to each sample. However, an average could only be computed for samples from studies that included at least 1 of the 10 frequently used tests. The decision to use the average eliminated four of the 302 samples in this review. Restricting the computation of average ES_{RM} to those 10 tests eliminated 10.0% (71 of 713) of the ES_{RM}s and 9.6% (1,001 of 10,407) of the individual test scores. These losses were balanced by the fact that the mean ES for every remaining test was based on a reasonably broad ES_{RM} sample and a substantial minimum cumulative sample size of 252 test scores.

Summary observations. Three comments are in order regarding the decision to use the average ES_{RM} to compute variance estimates for the analyses reported in the main body of this paper. First, the initial observations that ES_{RM} varied as a function of strength test could be explained by confounding of test type with demographic characteristics. If demographics were the basis for the original confounding, ES_{RM} estimates that are independent of demographics are the appropriate values to include in the analyses. Second, averaging effects within studies eliminates an important statistical problem. When multiple ES estimates are derived from a single sample, those estimates are not independent. Analyses should allow for the lack of independence (Gleser & Olkin, 1994). Using an average ES is a common method of dealing with this problem. Ideally, there should be some empirical justification for averaging. Table B3 provided this justification for the present purposes. The average ES_{RM} was not significantly different controlling for demographic differences. Third, it can be argued that all of the strength tests are indicators of a single general strength construct (Vickers, 2003). If all of the tests measure the same construct, the average effect is the best available estimate of the program impact on the basic underlying strength construct. Thus, averaging the ES_{RM} values for different strength tests administered to a particular sample provides an empirically, statistically, and conceptually defensible index of resistance training effects.

Analysis Weights

The decision to average ES_{RM} values affected the computation of the weights for the moderator analyses. Three different sets of weights were used at various points in the analyses. The first weight set was test-specific. Weights were computed using the unweighted test-specific average ES_{RM} as $d_{RM(Test)}$. For example, the average ES_{RM} for bench press was used to compute bench press weights, the average ES_{RM} for biceps curls was used to compute biceps curl weights, and so forth. This weight set was employed in the initial bivariate assessments of moderator effects.

The second weight set was computed after the analyses showed that test results could be reduced to a single ES_{RM} estimate by averaging the values for the 10 most frequently used tests. Average values were computed for each sample in the analysis. The average was based on one to eight tests. The unweighted mean value for the average (ES_{RM}, $d_{RM} = 1.81$) was used in these computations.

The third weight set allowed for the effect of averaging within samples. An average score should be a more accurate estimate of the training effect than an individual score. Averaging ES_{RM} was analogous to the effect of averaging several items in a questionnaire to obtain an individual's scale score. If the average is derived from independent observations, the variance of the mean is the sum of the individual variances divided by the square of the number of observations. The third weight set was computed to obtain the variance associated with the second weight set by k^2 where k was the number of effects that were being averaged for the sample. Note that this correction is conservative because it does not allow for the correlation of scores from different tests.

The inverse of the variances served as the weight variable for the moderator analyses. The third weight set was used in the analyses described in the main body of this report. All of the analyses had the average ES_{RM} or the adjusted average ES_{RM} as the dependent variable. The choice of which weight to use was of more importance conceptually than in practice. The three weight sets were highly correlated ($r \geq .964$).

Appendix C

Tucker-Lewis Index

The TLI (Tucker & Lewis, 1973) was introduced to guard against what may be a wide spread problem in meta-analysis. Moderator analyses begin with a significance test. If the null hypothesis is rejected, the moderator variable is accepted as a meaningful influence on ES even if the differences between groups are quite small.

Relying on significance tests to identify important results is a risky proposition in any statistical analysis. Statistical significance is the product of sample size and ES (Rosenthal & Rosnow, 1984). In meta-regression, ES might be labeled "meta-ES" because it reflects the differences in the primary ES across moderator groups. A significant meta-ES could indicate a substantial between-groups difference, but it does not rule out the possibility that small between-groups difference have been amplified by a large sample size. Although it follows that the meta-ES must be separated from sample size to properly interpret findings, this principle is not routinely applied to meta-analysis even though logic says it should be.

The TLI was adapted to provide a meta-ES metric. The TLI, which is the proportion of greater than chance variation in ES, can be computed from the χ^2 values from a moderator analysis. The variation in ES_{RM} determines the χ^2 values. The TLI equation is:

$$TLI = \frac{(\chi_{Null}{}^{2} / df_{Null}) - (\chi_{Model}{}^{2} / df_{Model})}{(\chi_{Null}{}^{2} / df_{Null}) - 1}. \tag{C1}$$

The expected value of χ^2/df ratio is 1, so the denominator of Equation C1 is the proportion of the observed variation in ES_{RM} that is greater than expected by chance. The numerator is the variation in ES_{RM} accounted for by the model (i.e., total ES_{RM} variation minus the residual ES_{RM} variation after fitting the model).

The TLI is a reasonable index of the meta-regression ES. This effect size index makes use of the noncentrality parameter (McDonald & Marsh, 1990) that is the basis for the *Q* statistic that provides the test for homogeneity of variance (Hedges, 1982). Thus, a meta-regression effect size based on TLI maintains a connection between effect size and the probability that a moderator will be statistically significant.

The TLI is not an exact parallel to the usual effect size indicators—such as the proportion of variance explained in an ANOVA. One reason is that TLI is analogous to Hays's (1963) ω^2 rather than the usual ε^2. The difference between the two is that the variance that would be expected by chance is subtracted from the variance explained when computing ω^2 but not when computing ε^2. This difference is the reason that TLI will be less than zero when $\chi_{Model}{}^{2} / df_{Model} > \chi_{Null}{}^{2} / df_{Null}$ because the numerator will be a negative number. This situation arises when the reduction in the χ^2 produced by a model is small relative to the number of parameters in the model. For this reason, the reported TLI is the value derived from Equation C1 or .00, whichever is larger.

The interpretation of the TLI employed Cohen's (1988) general criteria for ES evaluations. Cohen's criteria classify ESs on the basis of the proportion of observed variation explained by a predictor. In this case, TLI is the proportion of non-random variation in ES_{RM}, so Cohen's (1988) ES classification rule is a suitable index for characterizing the strength of association of moderator variables with ES_{RM}: small meta-ES, $.01 \leq TLI < .10$; moderate meta-ES, $.10 \leq TLI < .25$; large meta-ES, $TLI \geq .25$).

Appendix D

Analyses With Individual Effect Sizes as Independent Observations

An initial series of moderator analyses was undertaken with each ES_{RM} treated as an independent observation. This approach was adopted because it has been routinely used in other resistance training meta-analyses of simple ES. Its inclusion provided a reference point for assessing the impact of shifting to a repeated measures effect size. The weights used in this analysis were based on test-specific values for the average ES_{RM}.

Demographic characteristics generally had a stronger impact on individual ES_{RM} values than did program characteristics (see Table D1). The much larger χ^2 values in this analysis compared with the analysis of weighted average effects (see Table 3) was the most notable aspect of the findings. The impact of training status was particularly pronounced. The strength of the training status effect is illustrated by the fact that the χ^2 for this moderator was not much less than the sum of the χ^2s for all other moderators. Allowing for the degrees of freedom in each analysis, the choice between 1 set per session and 3 sets per session was the most important program determinant of the response.

Table D1

Based on Individual ES_{RM} Values

	Moderator		Residual	
	χ^2	*df*	χ^2	*df*
Demographics				
Gender	134.92	1	4096.18	532
Age group	111.42	1	5589.81	727
Training status	714.41	1	5375.78	692
Program elements				
Periodization	1.19	1	5491.45	712
Sessions per week	109.80	4	5514.89	710
Sets per session	199.89	5	3980.22	507
Repetitions per set	317.74	6	3385.66	432

A second analysis of the individual ESs was carried out with no weights. The analyses were simple ANOVA procedures with a moderator variable defining the groups in the analysis. The simple ESs in this analysis were computed by subtracting the pretest mean from the posttest mean and dividing by the pretest standard deviation. This analysis procedure appeared to replicate that which was used in meta-analyses by Rhea and his colleagues (Peterson et al., 2005; Rhea & Alderman, 2004; Rhea et al., 2002; Rhea et al., 2003).

Table D2

Analysis of Unweighted Simple ES

Moderator variable	No. ES	Avg ES	*F* test	Sig.	ε^2
Gender					
Men	362	1.36			
Women	162	1.93	36.69	.000	.037
Age					
Younger	497	1.16			
Older	222	1.38	84.29	.000	.072
Training status					
Untrained	512	1.67			
Trained	175	.77	105.15	.000	.090
Periodization					
Progressive	505	1.53			
Periodized	199	1.06	31.39	.000	.028
Sessions per week					
1	19	.58			
2	211	1.05			
3	448	1.66			
4	20	.69			
5	4	1.17	78.31	.000	.067
Sets per session					
1	87	.90			
2	29	.99			
3	330	1.79			
4	21	1.17			
5	22	1.09			
6	10	1.16	74.21	.000	.087
Repetitions per set					
<60%	15	2.27			
60%	16	1.63			
65%	9	1.77			
70%	17	1.21			
75%	177	1.11			
80%	127	1.95			
85%	63	1.62	64.45	.000	.089

Note. The average effect size was smaller because the computations did not allow for the pretest/posttest correlation. Most of the effects were highly significant ($p < .001$), but the average ES were uniformly in the small ES range.

The results of the unweighted analysis of individual ESs were broadly comparable to the results of the other analyses conducted in this study (see Table D2). The analyses indicated significantly larger ESs for women, older people, untrained individuals. The finding that non-periodized programs produced larger gains than periodized programs conflicted with Rhea and Alderman’s (2004) findings. ES did not display any simple pattern of associations to sessions per week, sets per session, or repetitions per set.

The analysis of unweighted simple ESs was extended to illustrate that the problems introduced by confounding of different moderator variables could yield misleading results in this type of analysis as well as in the analyses reported in the main body of this paper. A two-way analysis of variance was performed with training status and periodization as predictors of ES. Experience was strongly related to ES as it had been in the bivariate analyses ($F = 60.89$, $p < .001$, $\varepsilon^2 = .083$). The difference between untrained and trained weightlifters was largely unchanged (untrained, ES = 1.63; trained weightlifters, ES = .70). However, periodization was no longer a significant predictor of ES ($F = .80$, $p = .798$, $\varepsilon^2 = .001$). The difference between progressive and periodized programs was reduced from $d = 0.47$ to $d = 0.10$ (progressive, ES = 1.22; periodized, ES = 1.12). Based on this analysis, the confounding of periodization with experience explained the bivariate association of periodization with ES.

Appendix E

Comparison With Findings of Other Meta-Analytic Reviews

Research generates a body of evidence. By summarizing that evidence, meta-analyses indicate which findings should be taken as established facts. Discussions that establish the proper interpretation of those facts convert the evidence to reliable scientific knowledge (Ziman, 1978).

No single meta-analysis is likely to establish a definitive set of facts within a given research domain. Uncertainty is unavoidable because meta-analyses share some attributes with primary studies. The studies in a meta-analysis are only a sample from a universe of potential studies. The choice of characteristics to code, coding criteria, method of computing ES, and choice of analysis procedures are analogous to research measurements and analysis decisions that are made in primary studies. Considered in the abstract, meta-analyses have much in common with survey studies. As in survey studies, different reviews can produce different results because each review involves decisions at multiple choice points (Wanous, Malinek, & Sullivan, 1989) and choices will differ from one review to the next.

The parallels between the design and analysis of surveys and meta-analyses mean that developing a feeling for the robustness of findings from meta-analyses is important. This undertaking is loosely comparable to replicating survey findings. The study samples can be expected to overlap from one meta-analysis of a topic to the next, but the samples seldom are identical. The introduction of adjustments for program duration and repeated measures are examples of choices that can be seen as yielding a survey that involves replication with extensions of the earlier work. To develop the required feeling for how these modifications affected the meta-analytic results, the findings from this review were compared with the findings from reviews by Peterson et al. (2004), Rhea and Alderman (2004), Rhea et al. (2002); Rhea et al. (2003), and Wolfe et al. (2004). The present findings were not compared to the results of two prior reviews of the effectiveness of resistance training in children (Falk & Tenenbaum, 1996; Payne et al., 1997). The present findings could not be compared to those of Peterson, Rhea, and Alvar (2005) because that the findings of that review were only reported graphically.

Comparison Procedures

The comparisons with other reviews focused on the assessment of moderator effects. Three questions were posed for each moderator: Is the moderator effect statistically significant? Is the effect large enough to be important? Does the evidence identify a single best practice training option? The last question was equivalent to asking “Does the equivalence set contain a single option?” Affirmative answers to all three questions would identify a robust best practice.

Each comparison involved three analyses. First, the means and standard deviations reported in prior meta-analyses were reanalyzed. These reanalyses were needed to estimate the amount of variance explained by moderators and to perform post hoc comparisons to define equivalence sets. Second, the methods used in the comparison review were applied to the studies covered in this review. The unweighted ES was the dependent variable in an ANOVA that treated each test administered to a sample as an independent observation. Third, the ANOVAs were repeated with the individual ES_{RM} values replacing ES with weights based on test-specific average effects (see Appendix B). The senior author's last name has been used to label reanalyses of results from prior reviews. The analyses of the simple ES and ES_{RM} from this review have been labeled ANOVA and GLM (i.e., general linear model) to indicate the analysis methods.

Equivalence sets were defined in three steps. First, the statistical significance of the moderator effect was determined by ANOVA or GLM. Second, the variance explained by the moderator variable was computed. The effect size computations produced η^2 values for each moderator. This statistic was used for all moderator analyses. Third, post hoc comparisons were performed.

The post hoc comparisons began by ranking the cells defined by the moderator from largest average ES to smallest average ES. The difference between the highest average and the second highest average was tested for statistical significance. If the difference was statistically nonsignificant, the difference between the highest average and the third highest average was tested for significance. The sequence of tests continued until a significant difference was identified. Fisher's least significant difference was the significance criterion for the analyses of the simple ES outcome measures. A Bonferroni adjustment was used in the ES_{RM} analyses. The Bonferroni adjustment assumed that the moderator level with the highest average would be compared to each other level of the moderator. The equivalence sets consisted of the moderator level with the highest average ES plus the other levels that were not significantly different from this reference point.

Each data analysis from this review was performed after imposing selection criteria that matched the data as closely as possible to the reference review. For example, Peterson et al. (2004) limited their review to athletes. The corresponding analyses of the present data were limited to data from studies of competitive weightlifters or other types of athletes who routinely engage in resistance training (e.g., football players).

Rhea et al. (2003) Comparison – Dose Response for Strength Development

Rhea et al. (2003) provided the broadest picture of the evidence for any of the comparison reviews. Their review covered 140 studies that produced 1433 ESs. Preliminary analyses showed a significant difference in training response between trained and untrained individuals and no difference for men and women. Based on those findings, other potential moderators were examined separately for trained and untrained individuals with samples of men and women within each category (see Table E1). The corresponding analyses of the present data included a similar division into trained and untrained individuals with separate analyses for each group.

Table E1

Rhea et al. (2003) Comparison

	Moderator	Analysis	Significance	η^2	Equivalence set
Trained	Intensity	Rhea[a]	.000	.168*	{80%}
		ANOVA[b]	.001	.327*	{80%, 60%}
		GLM[c]	.011	.169*	{60%, 85%, 80%}
	Sets	Rhea	.187		{4, 5, 3, 2, 1}
		ANOVA	.399		{4, 3, 5, 1}
		GLM	.000	.162*	{4}
	Sessions	Rhea	.000	.100*	{2}[d]
		ANOVA	.004	.078*	{3}
		GLM	.731	.002	{3, 2, 4, 1}
Untrained					
	Intensity	Rhea	.000	.033	{60%, 75%, 40%, 80%, 50%}
		ANOVA	.000	.073	{<60%, 60%, 80%, 85%, 65%}
		GLM	.000	.093	{85%, <60%, 60%, 65%}
	Sets	Rhea	.000	.035	{4, 3, 2}
		ANOVA	.000	.086*	{3, 5}
		GLM	.000	.074*	{3, 5}
	Sessions	Rhea	.000	.014	{3}
		ANOVA	.013	.021	{3}
		GLM	.000	.019	{2}

[a] Reanalysis of Rhea et al. (2003) data. [b] ANOVA for individual ES from the present review. [c] ANOVA for aggregated ES from the present review. [d] Dichotomous comparison.
*Moderator effect exceeds Cohen's (1988) criterion for a small effect.

Table E2

Wolfe et al. (2004) Comparison

		η^2		
Moderator	Analysis	Sets	Moderator	S x M interaction
Age	Wolfe	.000	.206**	.031
	ANOVA	.038	.004	.016**
	GLM	.033	.000	.001
Length	Wolfe	.008	.021	.082**
	ANOVA	.032**	.008*	.000
	GLM	.035**	.005**	.003
Men vs. women	Wolfe	.004	.010#	.011
	ANOVA	.000	.027*	.001
	GLM	.006	.002	.000
Training status	Wolfe	.002	.177**	.004
	ANOVA	.011*	.035**	.005
	GLM	.012	.052**	.001

Note. # p = .099; Cohen's (1988) criteria would classify every significant difference as representing a small, but potentially important effect.
*p<.05. **p < .01

Statistical significance. Intensity was the only moderator that was consistently significant for trained individuals. All three moderators were consistently significant for untrained individuals.

Variance explained. The intensity effect for trained individuals was the only moderator that consistently exceeded Cohen's (1988) minimum criterion for a small effect.

Equivalence sets. There was no equivalence set that was consistent across the three analyses.

Wolfe et al. (2004) Comparison – Single Versus Multiple Sets

Wolfe et al. (2004) reported the results of a review that covered 16 studies with 103 ESs (see Table E2). Means and standard deviations were reported for two-way classifications. Single set versus multiple sets was a factor in each classification. The other factors were gender, age, training status, and program length. Wolfe et al. (2004) reported the mean and standard deviation for each cell defined by each two-way classification. The cells in each cross-classification were analyzed as a one-way ANOVA. The reanalysis reconfigured the cells within each classification as a two-way ANOVA. Wolfe et al. (2004) did not indicate any selection criteria, so the reanalyses were carried out using all of the ESs in the present data.

Table E3

Rhea et al. (2002) Comparison: Single Versus Multiple Sets Review

	Moderator	Significance	η^2	Equivalence set
Training status	Rhea	.061	.062*	{Trained, untrained}
	ANOVA	.000	.090*	{Untrained}
	GLM	.000	.133*	{Untrained}
Length	Rhea	.287	.027	{11-15, 21-25, 6-10}
(in weeks)	ANOVA	.000	.045	{21 – 25}
	GLM	.000	.035	{21 – 25}

*Exceeds Cohen’s (1988) criterion for classification as a small effect.

Table E2 reports η^2 values for the main effects and the interaction effect. The η^2 values are based on the variance explained after controlling for the other effects in the table (i.e., unique sums of squares). Because each main effect and interaction involved a single degree of freedom, any $\eta^2 > .010$ met Cohen’s (1988) minimum criterion for a potentially important influence on the training response.

Training status was the only consistent effect. Although not shown in the table, the average values indicated that the training effect was much stronger for untrained than trained individuals

Rhea et al. (2002) – Multiple Sets

Rhea et al. (2002) provided a different approach to comparing single and multiple set programs (see Table E3). A set of 16 relevant studies produced 93 ESs. The ES computations treated multiple sets as the experimental group and single sets as the control group. The ES was based on the difference between the changes observed in the two conditions. This review was the only one that based ES on the difference in the improvements produced by two training programs. The current data set included some studies for which a single set program could have been matched to a multiple set program in the same study. However, this matching was not undertaken. Instead, the analyses of the present data approximated the Rhea et al. (2002) analysis by using two-way ANOVAs to estimate the effects of training status and program length controlling for single versus multiple sets.

Statistical significance. Training status was a consistent finding even though it was only marginally significant in the present reanalysis of Rhea et al.'s (2003) data (see Table E3). The effect of training status just reached statistical significance in the original analyses, $F_{1,55} = 4.03$, $p < .0497$. The difference may be the result of rounding the means and standard deviations to two decimal places when the initial findings were reported.

Variance explained. Training status met Cohen's (1988) minimum criterion for a small ES.

Equivalence sets. Untrained individuals consistently displayed greater training effects than trained individuals.

Rhea and Alderman (2004) Comparison

Rhea and Alderman (2004) reviewed the effects of periodized resistance training programs. Their review covered data from 105 studies that produced at least 649 ESs for the moderator analyses.

Rhea and Alderman (2004) performed a test of the overall difference between periodized and non-periodized programs as the first step in their analysis (see Table E4). When that comparison indicated statistically significant differences, they chose to include only the effects from periodized programs in subsequent moderator analyses of moderator variables.

The difference between periodized and non-periodized programs was statistically significant and large enough to be important in Rhea and Alderman's (2004) analysis and when the same methods were applied to the present data. However, the periodization effect was not even statistically significant in the GLM analysis. This inconsistency suggested that the choice of analysis procedures was an important factor in these analyses. Subsequent moderator tests were limited to just those ESs from groups that completed periodized programs.

Statistical significance. Gender, age, training status, and program length were consistently significant moderators of ES.

Variance explained. Gender and training status consistently met the Cohen criterion.

Equivalence sets. Untrained individuals consistently displayed stronger training effects.

Table E4

Rhea and Alderman (2004) Comparison: Meta-Analyses for Periodization

	Review	Significance	η^2	Equivalence set
Overall	Rhea	.000	.013*	{Periodized, non-periodized}[a]
	ANOVA	.000	.028*	{Non-periodized}
	GLM	.275	.000	{Non-periodized, periodized}
Within periodized programs				
Gender	Rhea	.000	.132*	{Combined}[b]
	ANOVA	.000	.095*	{Women}
	GLM	.000	.136*	{Women}
Age (in years)	Rhea	.000	.020	{<55}[c]
	ANOVA	.050	.019	{>55}
	GLM	.029	.003	{>55, <55}
Training status	Rhea	.000	.110*	{Untrained}
	ANOVA	.000	.115*	{Untrained}
	GLM	.000	.336*	{Untrained}
Program length	Rhea	.019	.012	{9-20, 1-8, 20-40 weeks}
	ANOVA	.000	.142*	21-25 weeks
	GLM	.000	.055	{21-25, 26 – 40 weeks}

[a] Periodized would have been the only choice if the group labeled "Overall" had been omitted. [b] Men and women combined into a single sample. Men and women would have been assigned to a single group if the analysis had been limited to those two groups. [c] Moderator variable was a dichotomy.

Peterson et al. (2004) – Dose Response Relationship for Athletes

Peterson et al. (2004) examined the dose response relationship for athletes only. The review covered 37 studies with 370 ESs (see Table E5). The closest comparison group in this analysis was the trained weightlifters.

Table E5

Comparison of Dose Response Analyses for Athletes

	Analysis	Significance	η^2	Equivalence set
Intensity	Peterson	.000	.143*	{85%, 75%, 80%}
	ANOVA	.001	.327*	{80%, 60%)
	GLM	.000	.184*	{<60%, 85%, 75%}
Sets	Peterson	.027	.053	{8, 14, 4, 12, 6, 5, 16, 3, 1}
	ANOVA	.399	.044	{4, 3, 5, 1}
	GLM	.042	.062*	(1, 3, 4, 5}
Sessions	Peterson	.926	.000	{2, 3}
	ANOVA	.004	.078*	{3}
	GLM	.000	.099	{3,2}

Statistical significance. Intensity was the only consistently significant moderator.

Variance explained. Intensity consistently met Cohen’s (1988) minimum ES criterion.

Equivalence sets. None of the equivalence sets were identical. In fact, there was not a single intensity that was included in all 3 sets.

Comparison Summary

The three questions that guided the comparisons provided a sequential screening process to identify best practices. A total of 24 comparisons were made if the interaction terms for the Wolfe et al. (2004) comparison are excluded from consideration. Ten of 24 moderator effects were significant in all three analyses. Only five of the 10 consistently significant moderator effects also explained enough variance to meet Cohen’s (1988) minimum criterion for a small effect size. In three of the 10 cases, the variance explained was consistently less than Cohen’s (1988) minimum criterion. Finally, only two of the five cases that met the first two criteria consistently produced the same equivalence set. In both cases, the consistent equivalence set consisted solely of untrained individuals. Given that training status is not ordinarily thought of as an integral part of the program design, there was not a single instance of a best practice as the term has been defined in this report. It apparently has been easy to obtain a statistically significant moderator effect, but hard to move from there to the identification of a single best design option when variance explained and statistical significance criteria are applied before designating the nominal best practice.

References

*Article provided data for the meta-analysis.

*Abe, T., De Hoyos, D. V., Pollock, M. L., & Garzarella, L. (2000). Time course for strength and muscle thickness changes following upper and lower body resistance training in men and women. *Eur J Appl Physiol*, *81*(3), 174-180.

*Abernethy, P. J., & Jurimae, J. (1996). Cross-sectional and longitudinal uses of isoinertial, isometric, and isokinetic dynamometry. *Medicine and Science in Sports and Exercise, 28*(9), 1180-1187.

*Almasbakk, B., & Hoff, J. (1996). Coordination, the determinant of velocity specificity? *J Appl Physiol, 81*(5), 2046-2052.

American College of Sports Medicine position stand. (2009). Progression models in resistance training for healthy adults. *Med Sci Sports Exerc*, *41*(3), 687-708.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

*Anderson, T., & Kearney, J. T. (1982). Effects of three resistance training programs on muscular strength and absolute and relative endurance. *Research Quarterly for Exercise and Sport*, *53*, 1-7.

*Arce Esquivel, A. A., & Welsch, M. A. (2007). High and low volume resistance training and vascular function. *Int J Sports Med, 28*(3), 217-221.

*Baker, D. (2001). The effects of an in-season of concurrent training on the maintenance or maximal strength and power in professional and college-aged rugby league football players. *Journal of Strength and Conditioning Research*, *15*, 172-177.

*Baker, D., Wilson, G., & Carlyon, B. (1994). Generality versus specificity: A comparison of dynamic and isometric measures of strength and speed-strength. *Eur J Appl Physiol Occup Physiol, 68*(4), 350-355.

*Baker, D. G., & Newton, R. U. (2006). Adaptations in upper-body maximal strength and power output resulting from long-term resistance training in experienced strength-power athletes. *J Strength Cond Res, 20*(3), 541-546.

*Bartholomew, J. B., Stults-Kolehmainen, M. A., Elrod, C. C., & Todd, J. S. (2008). Strength gains after resistance training: The effect of stressful, negative life events. *J Strength Cond Res, 22*(4), 1215-1221.

*Becque, M. D., Lochmann, J. D., & Melrose, D. R. (2000). Effects of oral creatine supplementation on muscular strength and body composition. *Med Sci Sports Exerc*, *32*(3), 654-658.

*Bell, G. J., Syrotuik, D., Martin, T. P., Burnham, R., & Quinney, H. A. (2000). Effect of concurrent strength and endurance training on skeletal muscle properties and hormone concentrations in humans. *Eur J Appl Physiol, 81*(5), 418-427.

*Bell, G., Syrotuik, D., Socha, T., Maclean, I., & Quinney, H. A. (1997). Effect of strength training and concurrent strength and endurance training on strength, testosterone, and cortisol. *Journal of Strength and Conditioning Research*, *11*, 57-64.

*Bemben, M. G., Bemben, D. A., Loftiss, D. D., & Knehans, A. W. (2001). Creatine supplementation during resistance training in college football athletes. *Med Sci Sports Exerc*, *33*(10), 1667-1673.

*Berger, R. A. (1962). Effect of varied weight training programs on strength. *Res Q, 33,* 168-181.

*Blazevich, A. J., & Giorgi, A. (2001). Effect of testosterone administration and weight training on muscle architecture. *Med Sci Sports Exerc, 33*(10), 1688-1693.

*Blazevich, A. J., & Jenkins, D. G. (2002). Effect of the movement speed of resistance training exercises on sprint and strength performance in concurrently training elite junior sprinters. *J Sports Sci, 20*(12), 981-990.

*Bottaro, M., Machado, S. N., Nogueira, W., Scales, R., & Veloso, J. (2007). Effect of high- versus low-velocity resistance training on muscular fitness and functional performance in older men. *Eur J Appl Physiol, 99*(3), 257-264.

*Braith, R. W., Graves, J. E., Leggett, S. H., & Pollock, M. L. (1993). Effect of training on the relationship between maximal and submaximal strength. *Med Sci Sports Exerc, 25*(1), 132-138.

*Brandon, L. J., Sharon, B. F., Boyette, L. W., Anderson, K. A., & Stiles, R. (1997). Effects of resistive training on strength gains in older adults. *J Nutr Health Aging, 1*(2), 114-119.

*Broeder, C. E., Burrhus, K. A., Svanevik, L. S., & Wilmore, J. H. (1992). The effects of either high-intensity resistance or endurance training on resting metabolic rate. *Am J Clin Nutr, 55,* 802-810.

*Buford, T. W., Rossi, S. J., Smith, D. B., & Warren, A. J. (2007). A comparison of periodization models during nine weeks with equated volume and intensity for strength. *J Strength Cond Res, 21*(4), 1245-1250.

*Burke, D. G., Chilibeck, P. D., Parise, G., Candow, D. G., Mahoney, D., & Tarnopolsky, M. (2003). Effect of creatine and weight training on muscle creatine and performance in vegetarians. *Med Sci Sports Exerc, 35*(11), 1946-1955.

*Cacchio, A., Don, R., Ranavolo, A., Guerra, E., McCaw, S. T., Procaccianti, R., et al. (2008). Effects of 8-week strength training with two models of chest press machines on muscular activity pattern and strength. *J Electromyogr Kinesiol, 18*(4), 618-627.

*Candow, D. G., Little, J. P., Chilibeck, P. D., Abeysekara, S., Zello, G. A., Kazachkov, M. … Yu, P. H. (2008). Low-dose creatine combined with protein during resistance training in older men. *Med Sci Sports Exerc.*

*Cardoso Marques, M. A., & Gonzalez-Badillo, J. J. (2006). In-season resistance training and detraining in professional team handball players. *J Strength Cond Res, 20*(3), 563-571.

Carpinelli, R. N., & Otto, R. M. (1998). Strength training: Single versus multiple sets. *Sports Medicine, 26*(2), 73-84.

*Carroll, T. J., Abernethy, P. J., Logan, P. A., Barber, M., & McEniery, M. T. (1998). Resistance training frequency: Strength and myosin heavy chain responses to two and three bouts per week. *Eur J Appl Physiol Occup Physiol, 78*(3), 270-275.

*Channell, B. T., & Barfield, J. P. (2008). Effect of Olympic and traditional resistance training on vertical jump improvement in high school boys. *J Strength Cond Res, 22*(5), 1522-1527.

*Charette, S. L., McEvoy, L., Pyka, G., Snow-Harter, C., Guido, D., Wiswell *Strength Cond Res, 22*(5), 1522-1527., R. A., & Marcus, R. (1991). Muscle hypertrophy response to resistance training in older women. *J Appl Physiol, 70*(5), 1912-1916.

*Chilibeck, P. D., Calder, A. W., Sale, D. G., & Webber, C. E. (1998). A comparison of strength and muscle mass increases during resistance training in young women. *European Journal of Applied Physiology and Occupational Physiology, 77*, 170-175.

*Chromiak, J. A., Smedley, B., Carpenter, W., Brown, R., Koh, Y. S., Lamberth, J. G. … Altorfer, G. (2004). Effect of a 10-week strength training program and recovery drink on body composition, muscular strength and endurance, and anaerobic power and capacity. *Nutrition, 20*(5), 420-427.

*Coburn, J. W., Housh, D. J., Housh, T. J., Malek, M. H., Beck, T. W., Cramer, J. T. … Donlin, P. E. (2006). Effects of leucine and whey protein supplementation during eight weeks of unilateral resistance training. *J Strength Cond Res, 20*(2), 284-291.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second ed.). Hillsdale, NJ: Erlbaum.

*Cortez-Cooper, M. Y., Anton, M. M., DeVan, A. E., Neidre, D. B., Cook, J. N., & Tanaka, H. (2008). The effects of strength training on central arterial compliance in middle-aged and older adults. *Eur J Cardiovasc Prev Rehabil, 15*(2), 149-155.

*Cortez-Cooper, M. Y., DeVan, A. E., Anton, M. M., Farrar, R. P., Beckwith, K. A., Todd, J. S. & Tanaka, H. (2005). Effects of high intensity resistance training on arterial stiffness and wave reflection in women. *Am J Hypertens, 18*(7), 930-934.

*Cowley, P. M., Swensen, T., & Sforzo, G. A. (2007). Efficacy of instability resistance training. *Int J Sports Med, 28*(10), 829-835.

*De Hoyos, D. V., Abe, T., Garzarella, L., Hass, C., Nordman, M., & Pollock, M. (1998). Effects of 6 months of high- or low-volume resistance training on muscular strength and endurance. *Medicine and Science in Sports and Exercise, 30*, S165.

*DeBeliso, M., Harris, C., Spitzer-Gibson, T., & Adams, K. J. (2005). A comparison of periodised and fixed repetition training protocol on strength in older adults. *J Sci Med Sport, 8*(2), 190-199.

*Dolezal, B. A., & Potteiger, J. A. (1998). Concurrent resistance and endurance training influence basal metabolic rate in nondieting individuals. *J Appl Physiol, 85*(2), 695-700.

*Dorgo, S., King, G. A., & Rice, C. A. (2009). The effects of manual resistance training on improving muscular strength and endurance. *J Strength Cond Res, 23*(1), 293-303.

*Dornemann, T. M., McMurray, R. G., Renner, J. B., & Anderson, J. J. (1997). Effects of high-intensity resistance exercise on bone mineral density and muscle strength of 40-50-year-old women. *J Sports Med Phys Fitness, 37*(4), 246-251.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

*Epperson, W. L., Burton, R. R., & Bernauer, E. M. (1985). The effectiveness of specific weight training regimens on simulated aerial combat maneuvering G tolerance. *Aviat Space Environ Med, 56*(6), 534-539.

*Fahey, T. D., & Brown, C. H. (1973). The effects of an anabolic steroid on the strength, body composition, and endurance of college males when accompanied by a weight training program. *Med Sci Sports, 5*(4), 272-276.

Falk, B., & Tenenbaum, G. (1996). The effectiveness of resistance training in children. A meta-analysis. *Sports Med, 22*(3), 176-186.

*Falk, D. J., Heelan, K. A., Thyfault, J. P., & Koch, A. J. (2003). Effects of effervescent creatine, ribose, and glutamine supplementation on muscular strength, muscular endurance, and body composition. *J Strength Cond Res*, *17*(4), 810-816.

*Fatouros, I. G., Jamurtas, A. Z., Leontsini, D., Taxildaris, K., Aggelousis, N., Kostopoulus, N. … Buckemeyer, P. (2000). Evaluation of plyometric exercise training, weight training, and their combination on vertical jumping performance and leg strength. *Journal of Strength and Conditioning Research, 14*, 470-476.

*Fatouros, I. G., Kambas, A., Katrabasas, I., Leontsini, D., Chatzinikolaou, A., Jamurtas, A. Z. …. Taxildaris, K. (2006). Resistance training and detraining effects on flexibility performance in the elderly are intensity-dependent. *J Strength Cond Res*, *20*(3), 634-642.

*Fatouros, I. G., Kambas, A., Katrabasas, I., Nikolaidis, K., Chatzinikolaou, A., Leontsini, D., & Taxildaris, K. (2005). Strength training and detraining effects on muscular strength, anaerobic power, and mobility of inactive older men are intensity dependent. *Br J Sports Med*, *39*(10), 776-780.

*Fatouros, I. G., Tournis, S., Leontsini, D., Jamurtas, A. Z., Sxina, M., Thomakos, P … Mitrakous, A. (2005). Leptin and adiponectin responses in overweight inactive elderly following resistance training and detraining are intensity related. *J Clin Endocrinol Metab*, *90*(11), 5970-5977.

*Fawcett, J. P., Farquhar, S. J., Walker, R. J., Thou, T., Lowe, G., & Goulding, A. (1996). The effect of oral vanadyl sulfate on body composition and performance in weight-training athletes. *Int J Sport Nutr*, *6*(4), 382-390.

*Flynn, M. G., Fahlman, M., Braun, W. A., Lambert, C. P., Bouillon, L. E., Brolinson, P. G., et al. (1999). Effects of resistance training on selected indexes of immune function in elderly women. *J Appl Physiol*, *86*(6), 1905-1913.

*Folland, J. P., Chong, J., Copeman, E. M., & Jones, D. A. (2001). Acute muscle damage as a stimulus for training-induced gains in strength. *Med Sci Sports Exerc*, *33*(7), 1200-1205.

*Frontera, W. R., Meredith, C. N., O'Reilly, K. P., & Evans, W. J. (1990). Strength training and determinants of VO2max in older men. *J Appl Physiol*, *68*(1), 329-333.

*Frontera, W. R., Meredith, C. N., O'Reilly, K. P., Knuttgen, H. G., & Evans, W. J. (1988). Strength conditioning in older men: Skeletal muscle hypertrophy and improved function. *J Appl Physiol*, *64*(3), 1038-1044.

*Fry, A. C., Bonner, E., Lewis, D. L., Johnson, R. L., Stone, M. H., & Kraemer, W. J. (1997). The effects of gamma-oryzanol supplementation during resistance exercise training. *Int J Sport Nutr*, *7*(4), 318-329.

Galvao, D. A., & Taaffe, D. R. (2004). Single- vs. multiple-set resistance training: Recent developments in the controversy. *Journal of Strength and Conditioning Research*, *18*(3), 660-667.

*Galvao, D. A., & Taaffe, D. R. (2005). Resistance exercise dosage in older adults: Single- versus multiset effects on physical performance and body composition. *J Am Geriatr Soc*, *53*(12), 2090-2097.

*Garcia-Lopez, D., Hakkinen, K., Cuevas, M. J., Lima, E., Kauhanen, A., Mattila, M.…Gonzalez-Gallego, J. (2007). Effects of strength and endurance training on antioxidant enzyme gene expression and activity in middle-aged men. *Scand J Med Sci Sports, 17*(5), 595-604.

*Gearhart, R. F., Jr., Lagally, K. M., Riechman, S. E., Andrews, R. D., & Robertson, R. J. (2008). RPE at relative intensities after 12 weeks of resistance-exercise training by older adults. *Percept Mot Skills, 106*(3), 893-903.

*Gettman, L. R., Ward, P., & Hagan, R. D. (1982). A comparison of combined running and weight training with circuit weight training. *Med Sci Sports Exerc, 14*, 229-234.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *J Educ Stat, 18*, 271-279.

*Giorgi, A., Weatherby, R. P., & Murphy, P. W. (1999). Muscular strength, body composition and health responses to the use of testosterone enanthate: A double blind study. *J Sci Med Sport, 2*(4), 341-355.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *Handbook of Research Synthesis* (pp. 339-355). New York: Russell Sage Foundation.

Hagerman, P. S. (2001). *Comparison of a strength-training program performed on the Concept2 Dyno and Cybex chest press machine among college age men.* Unpublished Doctoral Dissertation, Oklahoma State University, Stillwater, OK.

*Hakkinen, K., Pakarinen, A., Kraemer, W. J., Hakkinen, A., Valkeinen, H., & Alen, M. (2001). Selective muscle hypertrophy, changes in EMG and force, and serum hormones during strength training in older women. *J Appl Physiol, 91*(2), 569-580.

*Harber, M. P., Fry, A. C., Rubin, M. R., Smith, J. C., & Weiss, L. W. (2004). Skeletal muscle and hormonal adaptations to circuit weight training in untrained men. *Scand J Med Sci Sports, 14*(3), 176-185.

*Harris, G. R., Stone, M. E., O'Bryant, H. S., Proulx, C. M., & Johnson, R. L. (2000). Short-term performance effects of high power, high force, or combined weight-training methods. *J Strength Cond Res, 14*, 14-20.

*Harris, K. A., & Holly, R. G. (1987). Physiological response to circuit weight training in borderline hypertensive subjects. *Med Sci Sports Exerc, 19*, 246-252.

*Hartman, J. W., Tang, J. E., Wilkinson, S. B., Tarnopolsky, M. A., Lawrence, R. L., Fullerton, A. V., et al. (2007). Consumption of fat-free fluid milk after resistance exercise promotes greater lean mass accretion than does consumption of soy or carbohydrate in young, untrained, male weightlifters. *Am J Clin Nutr, 86*(2), 373-381.

*Hartman, M. J., Clark, B., Bembens, D. A., Kilgore, J. L., & Bemben, M. G. (2007). Comparisons between twice-daily and once-daily training sessions in male weight lifters. *Int J Sports Physiol Perform, 2*(2), 159-169.

*Hass, C. J., Garzarella, L., De Hoyos, D., & Pollock, M. L. (2000). Single versus multiple sets in long-term recreational weightlifters. *Med Sci Sports Exerc, 32*(1), 235-242.

*Hasten, D. L., Rome, E. P., Franks, B. D., & Hegsted, M. (1992). Effects of chromium picolinate on beginning weight training students. *Int J Sport Nutr, 2*(4), 343-350.

Hays, W. L. (1963). *Statistics for Psychologists*. New York: Holt, Rinehart, Winston.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychol Bull*, *92*, 490-499.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

*Hickson, R. C., Hidaka, K., & Foster, C. (1994). Skeletal muscle fiber type, resistance training, and strength-related performance. *Med Sci Sports Exerc, 26*(5), 593-598.

*Hickson, R. C., Hidaka, K., Foster, C., Falduto, M. T., & Chatterton, R. T., Jr. (1994). Successive time courses of strength development and steroid hormone responses to heavy-resistance training. *J Appl Physiol, 76*(2), 663-670.

*Hickson, R. C., Rosenkoetter, M. A., & Brown, M. M. (1980). Strength training effects on aerobic power and short-term endurance. *Med Sci Sports Exerc, 12*(5), 336-339.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research*, *11*, 325-344.

*Hoff, J., Helgerud, J., & Wisloff, U. (1999). Maximal strength training improves work economy in trained female cross-country skiers. *Med Sci Sports Exerc, 31*, 870-877.

*Hoffman, J. R., Cooper, J., Wendell, M., & Kang, J. (2004). Comparison of Olympic vs. traditional power lifting training programs in football players. *J Strength Cond Res, 18*(1), 129-135.

*Hoffman, J. R., & Klafeld, S. (1998). The effect of resistance training on injury rate and performance in a self-defense instructors course for women. *J Strength Cond Res, 12*, 52-56.

*Hoffman, J. R., Kraemer, W. J., Fry, A. C., Deschenes, M., & Kemp, M. (1990). The effects of self-selection for frequency of training in a winter conditioning program for football. *Journal of Applied Sport Science Research, 4*(3), 76-82.

*Hoffman, J. R., Ratamess, N. A., Cooper, J. J., Kang, J., Chilakos, A., & Faigenbaum, A. D. (2005). Comparison of loaded and unloaded jump squat training on strength/power performance in college football players. *J Strength Cond Res, 19*(4), 810-815.

*Hoffman, J. R., Ratamess, N. A., Kang, J., Falvo, M. J., & Faigenbaum, A. D. (2006). Effect of protein intake on strength, body composition and endocrine changes in strength/power athletes. *J Int Soc Sports Nutr, 3*, 12-18.

*Hoffman, J. R., Ratamess, N. A., Klatt, M., Faigenbaum, A. D., Ross, R. E., Tranchina, N. M.,…Kramer, W. J. (2009). Comparison between different off-season resistance training programs in Division III American college football players. *J Strength Cond Res, 23*(1), 11-19.

*Hoffman, J. R., Wendell, M., Cooper, J., & Kang, J. (2003). Comparison between linear and nonlinear in-season training programs in freshman football players. *J Strength Cond Res, 17*(3), 561-565.

*Horne, L., Bell, G., Fisher, B., Warren, S., & Janowska-Wieczorek, A. (1997). Interaction between cortisol and tumour necrosis factor with concurrent resistance and endurance training. *Clin J Sport Med, 7*(4), 247-251.

*Hortobagyi, T., & Katch, F. I. (1990). Role of concentric force in limiting improvement in muscular strength. *J Appl Physiol, 68*(2), 650-658.

*Hostler, D., Crill, M. T., Hagerman, F. C., & Staron, R. S. (2001). The effectiveness of 0.5-lb increments in progressive resistance exercise. *J Strength Cond Res, 15*(1), 86-91.

*Humburg, H., Baars, H., Schroder, J., Reer, R., & Braumann, K. M. (2007). 1-Set vs. 3-set resistance training: A crossover study. *J Strength Cond Res, 21*(2), 578-582.

Hunter, J. E. and F. L. Schmidt (1990). Methods of meta-analysis. Newbury Park, Sage Publications.

*Huso, M. E., Hampl, J. S., Johnston, C. S., & Swan, P. D. (2002). Creatine supplementation influences substrate utilization at rest. *J Appl Physiol, 93*(6), 2018-2022.

*Izquierdo, M., Hakkinen, K., Ibanez, J., Anton, A., Garrues, M., Ruesta, M., et al. (2003). Effects of strength training on submaximal and maximal endurance performance capacity in middle-aged and older men. *J Strength Cond Res, 17*(1), 129-139.

*Izquierdo, M., Hakkinen, K., Ibanez, J., Garrues, M., Anton, A., Zuniga, A., et al. (2001). Effects of strength training on muscle power and serum hormones in middle-aged and older men. *J Appl Physiol, 90*(4), 1497-1507.

*Jensen, J. L., Marstrand, P. C. D., & Nielsen, J. B. (2005). Motor skill training and strength training are associated with different plastic changes in the central nervous system. *J Appl Physiol, 99*, 1558-1568.

*Johnson, B. L., Adamczyk, J. W., Tennoe, K. O., & Stromme, S. B. (1976). A comparison of concentric and eccentric muscle training. *Med Sci Sports, 8*(1), 35-38.

*Jones, K., Bishop, P., Hunter, G., & Fleisig, G. (2001). The effects of varying resistance-training loads on intermediate- and high-velocity-specific adaptations. *J Strength Cond Res, 15*(3), 349-356.

*Jones, K., Hunter, G., Fleisig, G., Escamilla, R., & Lemak, L. (1999). The effects of compensatory acceleration on upper-body strength and power in collegiate football players. *Journal of Strength and Conditioning Research, 13*(2), 99-105.

*Jurimae, J., Abernethy, P. J., Blake, K., & McEniery, M. T. (1996). Changes in the myosin heavy chain isoform profile of the triceps brachii muscle following 12 weeks of resistance training. *Eur J Appl Physiol Occup Physiol, 74*(3), 287-292.

*Kalapotharakos, V. I., Michalopoulou, M., Godolias, G., Tokmakidis, S. P., Malliou, P. V., & Gourgoulis, V. (2004). The effects of high- and moderate-resistance training on muscle function in the elderly. *J Aging Phys Act, 12*(2), 131-143.

*Kalapotharakos, V., Smilios, I., Parlavatzas, A., & Tokmakidis, S. P. (2007). The effect of moderate resistance strength training and detraining on muscle strength and power in older men. *J Geriatr Phys Ther, 30*(3), 109-113.

*Kalapotharakos, V. I., Tokmakidis, S. P., Smilios, I., Michalopoulos, M., Gliatis, J., & Godolias, G. (2005). Resistance training in older women: Effect on vertical jump and functional performance. *J Sports Med Phys Fitness, 45*(4), 570-575.

Karabulut, M. (2008). *A comparison of the effects of two resistance training protocols on physiological responses in middle-aged men.* Unpublished Doctoral Dissertation, Oklahoma University, Norman, OK.

*Keeler, L. K., Finkelstein, L. H., Miller, W., & Fernhall, B. (2001). Early-phase adaptations of traditional-speed vs. super slow resistance training on strength and aerobic capacity in sedentary individuals. *J Strength Cond Res*, *15*(3), 309-314.

*Kerksick, C. M., Rasmussen, C. J., Lancaster, S. L., Magu, B., Smith, P., Melton, C. … Kreider, R.(2006). The effects of protein and amino acid supplementation on performance and training adaptations during ten weeks of resistance training. *J Strength Cond Res, 20*(3), 643-653.

*Kerr, D., Morton, A., Dick, I., & Prince, R. (1996). Exercise effects on bone mass in postmenopausal women are site-specific and load-dependent. *J Bone Miner Res, 11*(2), 218-225.

*Kirk, E. P., Washburn, R. A., Bailey, B. W., LeCheminant, J. D., & Donnelly, J. E. (2007). Six months of supervised high-intensity low-volume resistance training improves strength independent of changes in muscle mass in young overweight men. *J Strength Cond Res, 21*(1), 151-156.

Konstanty, P. G. (1990). *The effects of concentric and eccentric bench-press training on strength development.* Unpublished Master's Thesis, San Diego State University, San Diego, CA.

*Kotzamanidis, C., Chatzopoulos, D., Michailidis, C., Papaiakovou, G., & Patikas, D. (2005). The effect of a combined high-intensity strength and speed training program on the running and jumping ability of soccer players. *J Strength Cond Res*, *19*(2), 369-375.

*Kraemer, W. J. (1997). A series of studies: The physiological basis for strength training in American football: Fact over philosophy. *Journal of Strength and Conditioning Research, 11*, 131-142.

Kraemer, W. J., Adams, K., Cafarelli, E., Dudley, G. A., Dooly, C., Feigenbaum, M. S., et al. (2002). American College of Sports Medicine position stand. Progression models in resistance training for healthy adults. *Med Sci Sports Exerc, 34*(2), 364-380.

*Kraemer, W. J., Ratamess, N. A., Volek, J. S., Hakkinen, K., Rubin, M. R., French, D. N., et al. (2006). The effects of amino acid supplementation on hormonal responses to resistance training overreaching. *Metabolism*, *55*(3), 282-291.

*Kraemer, W. J., Volek, J. S., Clark, K. L., Gordon, S. E., Puhl, S. M., Koziris, L. P., et al. (1999). Influence of exercise training on physiological and performance changes with weight loss in men. *Med Sci Sports Exerc, 31*(9), 1320-1329.

*Kramer, J., Stone, M. E., O'Bryant, H. S., & al., e. (1997). Effect of single vs. multiple sets of weight training: Impact of volume, intensity, and variation. *J Strength Cond Res, 11*, 143-147.

*Landin, D., & Nelson, A. G. (2007). Early phase strength development: A four-week training comparison of different programs. *J Strength Cond Res, 21*(4), 1113-1116.

*Larson-Meyer, D. E., Hunter, G. R., Trowbridge, C. A., Turk, J. C., Ernest, J. M., Torman, S. L. & Harbin, P. A. (2000). The effect of creatine supplementation on muscle strength and body composition during off-season training in female soccer players. *J Strength Cond Res, 14*, 434-442.

*Lemmer, J. T., Martel, G. F., Hurlbut, D. E., & Hurley, B. F. (2007). Age and sex differentially affect regional changes in one repetition maximum strength. *J Strength Cond Res, 21*(3), 731-737.

*Lexell, J., Downham, D. Y., Larsson, Y., Bruhn, E., & Morsing, B. (1995). Heavy-resistance training in older Scandinavian men and women: Short- and long-term effects on arm and leg muscles. *Scand J Med Sci Sports, 5*(6), 329-341.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

*Liu, Y., Schlumberger, **A.,** Wirth, K., Schmidtbleicher, D., & Steinacker, J. M. (2002). Different effects on human skeletal myosin heavy chain isoform expression: Strength vs. combination training. *J Appl Physiol, 94*, 2282-2288.

Lucier, S. G. (1999). *Impact of creatine monohydrate when incorporated with a strength and conditioning program.* Unpublished Master’s Thesis, Southwest State University, Marshall, MN.

*Lukaski, H. C., Bolonchuk, W. W., Siders, W. A., & Milne, D. B. (1996). Chromium supplementation and resistance training: Effects on body composition, strength, and trace element status of men. *Am J Clin Nutr, 63*(6), 954-965.

Marcinik, E. J. (1986). SPARTEN: A total body fitness program for health and physical readiness. *U.S. Navy Med*, *77*(2), 12-15.

Marcinik, E. J., Hodgdon, J. A., Englund, C. E., & O’Brien, J. J. (1987). Changes in fitness and shipboard task performance following circuit weight training programs featuring continuous or interval running. *Eur J Appl Physiol Occup Physiol, 56*(2), 132-137.

Marcinik, E. J., Hodgdon, J. A., & Vickers, R. R., Jr. (1985). The effects of an augmented and the standard recruit physical training program on fitness parameters. *Aviat Space Environ Med, 56*(3), 204-207.

*Marcinik, E. J., Potts, J., Schlabach, G., Will, S., Dawson, P., & Hurley, B. F. (1991). Effects of strength training on lactate threshold and endurance performance. *Med Sci Sports Exerc, 23*(6), 739-743.

*Margonis, K., Fatouros, I. G., Jamurtas, A. Z., Nikolaidis, M. G., Douroudos, I., Chatzinikolaou, A. … Kouretas, D. (2007). Oxidative stress biomarkers responses to physical overtraining: Implications for diagnosis. *Free Radic Biol Med, 43*(6), 901-910.

*Massey, C. D., Vincent, J., Maneval, M., & Johnson, J. T. (2005). Influence of range of motion in resistance training in women: Early phase adaptations. *J Strength Cond Res, 19*(2), 409-411.

*Massey, C. D., Vincent, J., Maneval, M., Moore, M., & Johnson, J. T. (2004). An analysis of full range of motion vs. partial range of motion training in the development of strength in untrained men. *J Strength Cond Res, 18*(3), 518-521.

*Masuda, K., Choi, J. Y., Shimojo, H., & Katsuta, S. (1999). Maintenance of myoglobin concentration in human skeletal muscle after heavy resistance training. *European Journal of Applied Physiology, 79*, 347-352.

*Mayhew, J. L., Johnson, B. D., Lamonte, M. J., Lauber, D., & Kemmler, W. (2008). Accuracy of prediction equations for determining one repetition maximum bench press in women before and after resistance training. *J Strength Cond Res, 22*(5), 1570-1577.

*Mayhew, J. L., Ware, J. S., Johns, R. A., & Bemben, M. G. (1997). Changes in upper body power following heavy-resistance strength training in college men. *Int J Sports Med, 18*(7), 516-520.

*McBride, J. M., Blaak, J. B., & Triplett-McBride, T. (2003). Effect of resistance exercise volume and complexity on EMG, strength, and regional body composition. *Eur J Appl Physiol, 90*(5-6), 626-632.

*McCarthy, J. P., Bamman, M. M., Yelle, J. M., LeBlanc, A. D., Rowe, R. M., Greenisen, M. C. … Fortney, S. M.(1997). Resistance exercise training and the orthostatic response. *Eur J Appl Physiol, 76*, 32-40.

*McCurdy, K., Langford, G., Ernest, J., Jenkerson, D., & Doscher, M. (2009). Comparison of chain- and plate-loaded bench press training on strength, joint pain, and muscle soreness in Division II baseball players. *J Strength Cond Res, 23*(1), 187-195.

McDonald, R. P. and H. W. Marsh (1990). “Choosing a multivariate model: Noncentrality and goodness of fit.” Psychol Bull, (107): 247-255.

*McLester, J. R., Bishop, P., & Guilliams, M. E. (2000). Comparison of 1 day and 3 days per week of equal-volume resistance training in experienced students. *J Strength Cond Res, 14*, 273-281.

*Meredith, C. N., Frontera, W. R., O’Reilly, K. P., & Evans, W. J. (1992). Body composition in elderly men: Effect of dietary modification during strength training. *J Am Geriatr Soc, 40*(2), 155-162.

Morris, S. B. (2000). Distribution of the standardized mean effect size for meta-analysis on repeated measures. *Br J Math Stat Psychol, 53*, 17-29.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105-125.

*Moss, B. M., Refsnes, P. E., Abildgaard, A., Nicolaysen, K., & Jensen, J. (1997). Effects of maximal effort strength training with different loads on dynamic strength, cross-sectional area, load-power and load-velocity relationships. *Eur J Appl Physiol Occup Physiol, 75*(3), 193-199.

*Munn, J., Herbert, R. D., Hancock, M. J., & Gandevia, S. C. (2005). Resistance training for strength: Effect of number of sets and contraction speed. *Med Sci Sports Exerc, 37*(9), 1622-1626.

National Research Council. (1992). Combining information: Statistical issues and opportunities for research. Washington, DC, National Academy Press.

*Neils, C. M., Udermann, B. E., Brice, G. A., Winchester, J. B., & McGuigan, M. R. (2005). Influence of contraction velocity in untrained individuals over the initial early phase of resistance training. *J Strength Cond Res, 19*(4), 883-887.

*Newton, R. U., Kraemer, W. J., & Hakkinen, K. (1999). Effects of ballistic training on preseason preparation of elite volleyball players. *Medicine and Science in Sports and Exercise, 31*(2), 323-330.

*Nichols, J. F., Omizo, D. K., Peterson, K. K., & Nelson, K. P. (1993). Efficacy of heavy-resistance training for active women over sixty: Muscular strength, body composition, and program adherence. *J Am Geriatr Soc, 41*(3), 205-210.

*Nobrega, A. C., Paula, K. C., & Carvalho, A. C. (2005). Interaction between resistance training and flexibility training in healthy young adults. *J Strength Cond Res, 19*(4), 842-846.

*Noonan, D., Berg, K., Latin, R. W., Wagner, J. C., & Reimers, K. (1998). Effects of varying dosages of oral creatine relative to fat free body mass on strength and body composition. *J Strength Cond Res, 12*, 104-108.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory (3rd ed.)*. New York: McGraw-Hill.

Omizo, D. K. K. (1992). *Effects of weight training on muscle strength of active older women.* Unpublished Master's Thesis, San Diego State University, San Diego, CA.

*Orsatti, F. L., Nahas, E. A., Maesta, N., Nahas-Neto, J., & Burini, R. C. (2008). Plasma hormones, muscle mass and strength in resistance-trained postmenopausal women. *Maturitas, 59*(4), 394-404.

*Osteras, H., Helgerud, J., & Hoff, J. (2002). Maximal strength-training effects on force-velocity and force-power relationships explain increases in aerobic performance in humans. *Eur J Appl Physiol, 88*(3), 255-263.

*Ostrowski, K. J., Wilson, G. J., Weatherby, R., Murphy, P. W., & Lyttle, A. D. (1997). The effect of weight training volume on hormonal output and muscular size and function. *Journal of Strength and Conditioning Research, 11*, 148-154.

*Panton, L. B., Rathmacher, J. A., Baier, S., & Nissen, S. (2000). Nutritional supplementation of the leucine metabolite beta-hydroxy-beta-methylbutyrate (hmb) during resistance training. *Nutrition, 16*(9), 734-739.

*Parkhouse, W. S., Coupland, D. C., Li, C., & Vanderhoek, K. J. (2000). IGF-1 bioavailability is increased by resistance training in older women with low bone mineral density. *Mech Ageing Dev, 113*(2), 75-83.

Payne, V. G., Morrow, J. R., Jr., Johnson, L., & Dalton, S. N. (1997). Resistance training in children and youth: A meta-analysis. *Res Q Exerc Sport, 68*(1), 80-88.

*Pearson, D. G., Hamby, D. G., Russel, W., & Harris, T. (1999). Long-term effects of creatine monohydrate on strength and power. *J Strength Cond Res, 13*, 187-192.

*Peeters, B. M., Lantz, C. D., & Mayhew, J. L. (1999). Effect of oral creatine monohydrate and creatine phosphate supplementation on maximal strength indices, body composition, and blood pressure. *J Strength Cond Res, 13*, 3-9.

*Pescatello, L. S., Kelsey, B. K., Price, T. B., Seip, R. L., Angelopoulos, T. J., Clarkson, P. M. … Hoffman, E. P. (2007). The muscle strength and size response to upper arm, unilateral resistance training among adults who are overweight and obese. *J Strength Cond Res, 21*(2), 307-313.

*Peterson, M. D., Dodd, D. J., Alvar, B. A., Rhea, M. R., & Favre, M. (2008). Undulation training for development of hierarchical fitness and improved firefighter job performance. *J Strength Cond Res, 22*(5), 1683-1695.

Peterson, M. D., Rhea, M. R., & Alvar, B. A. (2004). Maximizing strength development in athletes: A meta-analysis to determine the dose-response relationship. *J Strength Cond Res, 18*(2), 377-382.

Peterson, M. D., Rhea, M. R., & Alvar, B. A. (2005). Applications of the dose-response for muscular strength development: A review of meta-analytic efficacy and reliability for designing training prescription. *J Strength Cond Res, 19*(4), 950-958.

*Pipes, T. V. (1977). Physiological responses of fire fighting recruits to high intensity training. *J Occup Med, 19*(2), 129-132.

*Pipes, T. V. (1978). Variable resistance versus constant resistance strength training in adult males. *Eur J Appl Physiol Occup Physiol, 39*(1), 27-35.

*Pollock, M. L., Carroll, J. F., Graves, J. E., Leggett, S. H., Braith, R. W., Limacher, M., & Hagberg, J.M. (1991). Injuries and adherence to walk/jog and resistance training programs in the elderly. *Med Sci Sports Exerc, 23*(10), 1194-1200.

*Prokopy, M. P., Ingersoll, C. D., Nordenschild, E., Katch, F. I., Gaesser, G. A., & Weltman, A. (2008). Closed-kinetic chain upper-body training improves throwing performance of NCAA Division I softball players. *J Strength Cond Res, 22*(6), 1790-1798.

*Pruitt, L. A., Taaffe, D. R., & Marcus, R. (1995). Effects of a one-year high-intensity versus low-intensity resistance training program on bone mineral density in older women. *J Bone Miner Res, 10*(11), 1788-1795.

*Pyka, G., Lindenberger, E., Charette, S., & Marcus, R. (1994). Muscle strength and fiber adaptations to a year-long resistance training program in elderly men and women. *J Gerontol, 49*(1), M22-27.

*Reaburn, P., Logan, P., & Mackinnon, L. T. (1997). Serum testosterone response to high-intensity resistance training in male veteran sprint runners. *J Strength Cond Res, 11*, 256-260.

*Reynolds, T. H., Frye, P. A., & Sforzo, G. A. (1997). Resistance training and the blood lactate response to resistance exercise in women. *J Strength Cond Res, 11*, 77-81.

Rhea, M. R., & Alderman, B. L. (2004). A meta-analysis of periodized versus nonperiodized strength and power training programs. *Res Q Exerc Sport, 75*(4), 413-422.

*Rhea, M. R., Alvar, B. A., Ball, S. D., & Burkett, L. N. (2002). Three sets of weight training superior to 1 set with equal intensity for eliciting strength. *J Strength Cond Res, 16*(4), 525-529.

Rhea, M. R., Alvar, B. A., & Burkett, L. N. (2002). Single versus multiple sets for strength: A meta-analysis to address the controversy. *Res Q Exerc Sport, 73*(4), 485-488.

Rhea, M. R., Alvar, B. A., Burkett, L. N., & Ball, S. D. (2003). A meta-analysis to determine the dose response for strength development. *Med Sci Sports Exerc, 35*(3), 456-464.

*Rhea, M. R., Ball, S. D., Phillips, W. T., & Burkett, L. N. (2002). A comparison of linear and daily undulating periodized programs with equated volume and intensity for strength. *J Strength Cond Res, 16*(2), 250-255.

*Rhea, M. R., Phillips, W. T., Burkett, L. N., Stone, W. J., Ball, S. D., Alvar, B. A., & Thomas, A. B. (2003). A comparison of linear and daily undulating periodized programs with equated volume and intensity for local muscular endurance. *J Strength Cond Res, 17*(1), 82-87.

*Rhodes, E. C., Martin, A. D., Taunton, J. E., Donnelly, M., Warren, J., & Elliot, J. (2000). Effects of one year of resistance training on the relation between muscular strength and bone density in elderly women. *Br J Sports Med, 34*(1), 18-22.

*Roberts, M. D., Iosia, M., Kerksick, C. M., Taylor, L. W., Campbell, B., Wilborn, C. D. … Kreider, R. B. (2007). Effects of arachidonic acid supplementation on training adaptations in resistance-trained males. *J Int Soc Sports Nutr, 4*, 21. Retrieved from http//www.jissn.com/content/4/1/21

*Ronnestad, B. R., Kvamme, N. H., Sunde, A., & Raastad, T. (2008). Short-term effects of strength and plyometric training on sprint and jump performance in professional soccer players. *J Strength Cond Res, 22*(3), 773-780.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*(1), 185-193.

Rosenthal, R. (1979). The “file-drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.

*Rozenek, R., Ward, P., Long, S., & Garhammer, J. (2002). Effects of high-calorie supplements on body composition and muscular strength following resistance training. *J Sports Med Phys Fitness, 42*(3), 340-347.

Safrit, M. J., Hooper, L. M., Ehlert, S. A., Costa, M. G., & Patterson, P. (1988). The validity generalization of distance run tests. *Can J Sports Sci, 13*(4), 188-196.

*Sanborn, K., Boros, R., Hruby, J., Schilling, B., O’Bryant, H. S., Johnson, R. L. … Stone, M. E. (2000). Short-term performance effects of weight training with multiple sets not to failure vs. single set to failure in women. *Journal of Strength and Conditioning Research, 14*, 328-331.

*Schlicht, J., Camaione, D. N., & Owen, S. V. (2001). Effect of intense strength training on standing balance, walking speed, and sit-to-stand performance in older adults. *J Gerontol A Biol Sci Med Sci, 56*, M281-M286.

*Schlumberger, A., Stec, J., & Schmidtbleicher, D. (2001). Single- vs. multiple-set strength training in women. *J Strength Cond Res, 15*(3), 284-289.

*Schoitz, M., Potteiger, J. A., Huntsinger, P. G., & Denmark, D. C. (1998). The short-term effects of periodized and constant-intensity training on body composition, strength, and performance. *J Strength Cond Res, 12*, 173-178.

*Schroeder, E. T., Hawkins, S. A., & Jaque, S. V. (2004). Musculoskeletal adaptations to 16 weeks of eccentric progressive resistance training in young women. *J Strength Cond Res, 18*(2), 227-235.

*Sedliak, M., Finni, T., Peltonen, J., & Hakkinen, K. (2008). Effect of time-of-day-specific strength training on maximum strength and EMG activity of the leg extensors in men. *J Sports Sci, 26*(10), 1005-1014.

*Sharman, M. J., Newton, R. U., Triplett-McBride, T., McGuigan, M. R., McBride, J. M., Hakkinen, A. … Kraemer, W. J. (2001). Changes in myosin heavy chain composition with heavy resistance training in 60- to 75-year-old men and women. *Eur J Appl Physiol, 84*(1-2), 127-132.

Shimano, T., Kraemer, W. J., Spiering, B. A., Volek, J. S., Hatfield, D. L., Silvestre, R. … Hakkinen, K.(2006). Relationship between the number of repetitions and selected percentages of one repetition maximum in free weight exercises in trained and untrained men. *Journal of Strength and Conditioning Research, 20*(4), 819-823.

*Slivka, D., Raue, U., Hollon, C., Minchev, K., & Trappe, S. (2008). Single muscle fiber adaptations to resistance training in old (>80 yr) men: Evidence for limited skeletal muscle plasticity. *Am J Physiol Regul Integr Comp Physiol, 295*(1), R273-280.

*Smith, K., Winegard, K., Hicks, A. L., & McCartney, N. (2003). Two years of resistance training in older men and women: The effects of three years of detraining on the retention of dynamic strength. *Can J Appl Physiol, 28*(3), 462-474.

*Snow-Harter, C., Bouxsein, M. L., Lewis, B. T., Carter, D. R., & Marcus, R. (1992). Effects of resistance and endurance exercise on bone mineral status of young women: A randomized exercise intervention trial. *J Bone Miner Res, 7*(7), 761-769.

*Stanforth, P. R., Painter, T. L., & Wilmore, J. H. (1992). Alternation in concentric strength consequent to Powercise and Universal Gym circuit training. Journal of Applied Sport Science Research, *6*, 152-157.

*Stone, M. H., Potteiger, J. A., Pierce, K. C., Proulx, C. M., O'Bryant, H. S., Johnson, R. L., & Stone, M. E. (2000). Comparison of the effects of three different weight-training programs on the one repetition maximum squat. *Journal of Strength and Conditioning Research, 14*(3), 332-337.

*Storen, O., Helgerud, J., Stoa, E. M., & Hoff, J. (2008). Maximal strength training improves running economy in distance runners. *Med Sci Sports Exerc, 40*(6), 1087-1092.

*Swensen, T., Mancuso, P., & Howley, E. T. (1993). The effect of moderate resistance weight training on peak arm aerobic power. *Int J Sports Med, 14*(1), 43-47.

*Taaffe, D. R., Henwood, T. R., Nalls, M. A., Walker, D. G., Lang, T. F., & Harris, T. B. (2009). Alterations in muscle attenuation following detraining and retraining in resistance-trained older adults. *Gerontology, 55*, 217-223.

Tang, J.-L., & Liu, J. L. Y. (2000). Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology, 53*, 477-484.

*Tanimoto, M., Sanada, K., Yamamoto, K., Kawano, H., Gando, Y., Tabata, I. … Miyachi, M. (2008). Effects of whole-body low-intensity resistance training with slow movement and tonic force generation on muscular size and strength in young men. *J Strength Cond Res, 22*(6), 1926-1938.

*Tarnopolsky, M., Zimmer, A., Paikin, J., Safdar, A., Aboud, A., Pearce, E. … Doherty, T. (2007). Creatine monohydrate and conjugated linoleic acid improve strength and body composition following resistance exercise in older adults. *PLoS ONE, 2*(10), e991.

*Terzis, G., Georgiadis, G., Stratakos, G., Vogiatzis, I., Kavouras, S., Manta, P. … Blomstrand, E.. (2008). Resistance exercise-induced increase in muscle mass correlates with p70S6 kinase phosphorylation in human subjects. *Eur J Appl Physiol, 102*(2), 145-152.

*Terzis, G., Stratakos, G., Manta, P., & Georgiadis, G. (2008). Throwing performance after resistance training and detraining. *J Strength Cond Res, 22*(4), 1198-1204.

*Tokeshi, S. A., Kraemer, W. J., Nindl, B. C., Gotshalk, L. A., Marx, J. O., Harman, F. S. … Hakkinen, K. (1998). Power and strength in women: Adaptations following six months of resistance training. *Medicine and Science in Sports and Exercise, 30*(5), S165.

*Tracy, B. L., Byrnes, W. C., & Enoka, R. M. (2004). Strength training reduces force fluctuations during anisometric contractions of the quadriceps femoris muscles in old adults. *J Appl Physiol, 96*(4), 1530-1540.

*Tracy, B. L., & Enoka, R. M. (2006). Steadiness training with light loads in the knee extensors of elderly adults. *Med Sci Sports Exerc, 38*(4), 735-745.

*Trappe, S., Williamson, D., & Godard, M. (2002). Maintenance of whole muscle strength and size following resistance training in older men. *J Gerontol A Biol Sci Med Sci, 57*(4), B138-143.

*Trappe, S., Williamson, D., Godard, M., Porter, D., Rowden, G., & Costill, D. (2000). Effect of resistance training on single muscle fiber contractile function in older men. *Journal of Applied Physiology, 89*, 143-153.

*Tricoli, V., Lamas, L., Carnevale, R., & Ugrinowitsch, C. (2005). Short-term effects on lower-body functional power development: Weightlifting vs. vertical jump training programs. *J Strength Cond Res, 19*(2), 433-437.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

*Vandenberghe, K., Goris, M., Van Hecke, P., Van Leemputte, M., Vangerven, L., & Hespel, P. (1997). Long-term creatine intake is beneficial to muscle performance during resistance training. *Journal of Applied Physiology, 83*(6), 2055-2063.

Van Oosbree, P. A. (1993). *Strength training effects on leg strength, leg density, and cardiorespiratory fitness of females.* Unpublished Doctoral Dissertation, Texas Woman’s University, Denton, TX.

*Venable, M. P., Collins, M. A., O’Bryant, H. S., Denegar, C. F., Sedivec, J. J., & Alon, G. (1991). Effect of supplemental electric stimulation on the development of strength, vertical jump, performance and power. Journal of Applied Sport Science Research, *5*, 139-143.

*Verdijk, L. B., Jonkers, R. A., Gleeson, B. G., Beelen, M., Meijer, K., Savelberg, H. H. … van Loon, L. J. (2009). Protein supplementation before and after exercise does not further augment skeletal muscle hypertrophy after resistance training in elderly men. *Am J Clin Nutr, 89*(2), 608-616.

Vickers, R. R., Jr. (2003). The measurement structure of strength (Technical Report No. 03-30). San Diego, CA: Naval Health Research Center.

*Villareal, D. T., & Holloszy, J. O. (2006). DHEA enhances effects of weight training on muscle mass and strength in elderly women and men. *Am J Physiol Endocrinol Metab, 291*(5), E1003-1008.

*Vivodtzev, I., Wuyam, B., Flore, P., & Levy, P. (2005). Changes in quadriceps twitch tension in response to resistance training in healthy sedentary subjects. *Muscle Nerve, 32*(3), 326-334.

*Volek, J. S., Duncan, N. D., Mazzetti, S. A., Staron, R. S., Putukian, M., Gomez, A. L. … Kraemer, W. J. (1999). Performance and muscle fiber adaptations to creatine supplementation and heavy resistance training. *Med Sci Sports Exerc, 31*, 1147-1156.

*Walker, K. S., Kambadur, R., Sharma, M., & Smith, H. K. (2004). Resistance training alters plasma myostatin but not IGF-1 in healthy men. *Med Sci Sports Exerc, 36*(5), 787-793.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74*, 259-264.

Wathen, D., Baechle, T. R., & Earle, R. W. (2000). Training variation: Periodization. In T. R. Baechle & R. W. Earle (eds.), *Essentials of Strength Training and Conditioning* (pp. 513 – 527). Champaign, IL: Human Kinetics.

*Weir, J. P., Housh, D. J., Housh, T. J., & Weir, L. L. (1997). The effect of unilateral concentric weight training and detraining on joint angle specificity, cross-training, and the bilateral deficit. *J Orthop Sports Phys Ther*, *25*(4), 264-270.

*Westcott, W. L., Winett, R. A., Anderson, E. S., Wojcik, J. R., Loud, R. L., Cleggett, E., & Glover, S. (2001). Effects of regular and slow speed resistance training on muscle strength. *J Sports Med Phys Fitness, 41*(2), 154-158.

*Wilborn, C. D., Kerksick, C. M., Campbell, B. I., Taylor, L. W., Marcello, B. M., Rasmussen, C. J. … Kreider, R. B. (2004). Effects of Zinc Magnesium Aspartate (ZMA) Supplementation on Training Adaptations and Markers of Anabolism and Catabolism. *J Int Soc Sports Nutr, 1*(2), 12-20.

*Wilborn, C. D., Taylor, L. W., Campbell, B. I., Kerksick, C., Rasmussen, C. J., Greenwood, M., & Kreider, R. B. (2006). Effects of methoxyisoflavone, ecdysterone, and sulfo-polysaccharide supplementation on training adaptations in resistance-trained males. *J Int Soc Sports Nutr, 3*, 19-27.

*Wilder, N., Deivert, R. G., Hagerman, F., & Gilders, R. (2001). The effects of low-dose creatine supplementation versus creatine loading in collegiate football players. *J Athl Train, 36*(2), 124-129.

*Wilder, N., Gilders, R., Hagerman, F., & Deivert, R. G. (2002). The effects of a 10-week, periodized, off-season resistance-training program and creatine supplementation among collegiate football players. *J Strength Cond Res, 16*(3), 343-352.

*Willardson, J. M., & Burkett, L. N. (2008). The effect of different rest intervals between sets on volume components and strength gains. *J Strength Cond Res, 22*(1), 146-152.

*Williamson, D. L., Godard, M. P., Porter, D. A., Costill, D. L., & Trappe, S. W. (2000). Progressive resistance training reduces myosin heavy chain coexpression in single muscle fibers from older men. *J Appl Physiol, 88*(2), 627-633.

*Willoughby, D. S., & Simpson, S. (1998). Supplemental EMS and dynamic weight training: Effects on knee extensor strength and vertical jump of female college track & field athletes. *Journal of Strength and Conditioning Research, 12*(3), 131-137.

*Wilmore, J. H. (1974). Alterations in strength, body composition and anthropometric measurements consequent to a 10-week weight training program. *Med Sci Sports, 6*(2), 133-138.

*Wilmore, J. H., Parr, R. B., Girandola, R. N., Ward, P., Vodak, P. A., Barstow, T. J. … Leslie, P. (1978). Physiological alterations consequent to circuit weight training. *Med Sci Sports, 10*(2), 79-84.

*Wilson, G. J., Murphy, A. J., & Walshe, A. (1996). The specificity of strength training: The effect of posture. *Eur J Appl Physiol Occup Physiol, 73*(3-4), 346-352.

Wolfe, B. L., LeMura, L. M., & Cole, P. J. (2004). Quantitative analysis of single- vs. multiple-set programs in resistance training. *J Strength Cond Res, 18*(1), 35-47.

Womer, B. (2003). *Effect of a high- versus low-repetition resistance training program on muscular function in women.* Unpublished Master's Thesis, Southern Connecticut State University, New Haven, CT.

*Woolstenhulme, M. T., Conlee, R. K., Drummond, M. J., Stites, A. W., & Parcell, A. C. (2006). Temporal response of desmin and dystrophin proteins to progressive resistance exercise in human skeletal muscle. *J Appl Physiol, 100*(6), 1876-1882.

*Yarrow, J. F., Borsa, P. A., Borst, S. E., Sitren, H. S., Stevens, B. R., & White, L. J. (2008). Early-phase neuroendocrine responses and strength adaptations following eccentric-enhanced resistance training. *J Strength Cond Res, 22*(4), 1205-1214.

*Yoshizawa, M., Maeda, S., Miyaki, A., Misono, M., Saito, Y., Tanabe, K. … Ajisaka, R. (2009). Effect of 12 weeks of moderate-intensity resistance training on arterial stiffness: A randomized controlled trial in women aged 32-59. *Br J Sports Med, 43*, 615-618.

Ziman, J. (1978). *Reliable knowledge: An exploration of the grounds for belief in science*. London: Cambridge University Press.

REPORT DOCUMENTATION PAGE

1. Report Date (DD MM YY)
18 02 10

2. Report Type
Technical Report

3. DATES COVERED (from - to)
Apr 09 – Sep 09

4. TITLE AND SUBTITLE
Resistance Training: Identifying best practices?

5a. Contract Number:
5b. Grant Number:
5c. Program Element:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number: 60704

6. AUTHORS
Ross R. Vickers, Jr., Amanda Barnard, & Linda K. Hervig

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Commanding Officer
Naval Health Research Center
140 Sylvester Rd.
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER
Report No. 10-21

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Commanding Officer
Naval Medical Research Center
503 Robert Grant Ave
Silver Spring, MD 20910-7500

Commander
Navy Medicine Support Command
P.O. Box 140
Jacksonville, FL 32213-0140

10. Sponsor/Monitor's Acronyms(s)
NMRC/NMSC

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

Resistance training increases muscle strength. Muscle strength gains are influenced by program design. This review attempted to identify design choices that would be best practices. A best practice is a design option that produces significantly better results than any other option. To ensure sensitive assessments of program design effects, statistical procedures adjusted for differences in program length, and allowed for the repeated measures structure of the study designs. Untrained individuals benefitted much more from training than trained individuals. Gender had little effect. Age effects differed for men and women. Given the impact of participant characteristics on the training response, the effects of different program design facets were examined separately for programs with untrained and trained participants. Periodization, number of sessions per week, number of sets per session, and intensity (number of repetitions per set) were significant moderators for untrained participants; sets per session and intensity were significant moderators for trained participants. However, comparisons generally showed that no single design option was significantly better than all others. The available evidence may rule out some design choices (e.g., a single set per session), but it is too limited to identify best practices.

14. SUBJECT TERMS
muscle strength, resistance training, meta-analysis

16. SECURITY CLASSIFICATION OF:

a. REPORT	**b.ABSTRACT**	**c. THIS PAGE**	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
UNCL	UNCL	UNCL	UNCL	65

18a. NAME OF RESPONSIBLE PERSON
Commanding Officer

18b. TELEPHONE NUMBER (INCLUDING AREA CODE)
COMM/DSN: (619) 553-8429

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18