A Meta-Analysis of Sex Differences in Activity Level

Lesley Reid Enns

THESIS

Submitted in partial fulfillment of the
requirements for the degree of Master of Arts
in Psychology in the Faculty of Graduate Studies
at the University of Manitoba, 1982.

Winnipeg, Manitoba

A META-ANALYSIS OF SEX DIFFERENCES IN ACTIVITY LEVEL

BY

LESLEY REID ENNS

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Acknowledgments

There are a number of people whom I would like to thank for their contributions to this thesis.

I want especially to express my gratitude to my advisor, Dr. Warren Eaton, for the support and encouragement he has given me in the process of completing this research. His enthusiasm, patience, and attention to detail exemplify for me what a fine teacher and researcher should be.

I would also like to thank my committee members, Dr. John Schallow and Dr. Lance Roberts, for their encouragement and helpful suggestions.

Finally, I would like to acknowledge the contribution of my husband, Ken, whose continuous understanding and support have sustained me through this and many other difficult projects.

Financial assistance came from SSHRC Research Grant 410-81-123 to Dr. Warren Eaton.

A Meta-Analysis of Sex Differences in Activity Level

Explanations for Sex Differences

Interest in the nature of the psychological differences between the sexes has never been stronger. The whole sex difference question is especially important, given that many social changes are based on the beliefs most of us have about the essential natures of men and women. It is necessary that these beliefs, about both the exact nature of the differences and their etiology, be as complete and as accurate as possible.

Two theories, the modeling/imitation theory and the differential socialization theory, are frequently given as explanations for the development of psychological sex differences in children. In The Psychology of Sex Differences, Maccoby and Jacklin (1974) describe the two theories and find that both are inadequate and incomplete.

The modeling/imitation explanation rests on the assumption that differential reinforcement alone cannot account for the rate and breadth of sex-role acquisition. According to this theory, children learn sex-typed behaviour by modeling themselves after adults. The problem is how to explain why male and female children learn different behaviours if both imitate the same models. Maccoby and Jacklin outline two different hypotheses to explain why a child imitates the same-sex parent. The first hypothesis is that the same-sex parent is more available. The model availability hypothesis is dismissed because of lack of evidence. The second hypothesis holds that children imitate the behaviour of the parent the child perceives to be most like himself, the same-sex parent. Maccoby and Jacklin point out this hypothesis is untenable because

children display sex-typed behaviour before they demonstrate any clear-cut preference for a same-sex model to imitate. Furthermore they find little research evidence to support the existence of within-sex parent-child similarities on any types of behaviour that have been measured. It seems that children do not necessarily imitate the behaviour of the same-sex parent, as even their stereotypic sex-role behaviour does not correlate with that of the same-sex parent. Consequently the model/imitation explanation for the development of sex differences is seriously weakened. Maccoby and Jacklin argue, therefore, that the acquisition of behaviour through modeling is not sex-typed although the performance of behaviour might be. In other words, boys and girls learn masculine and feminine behaviour through modeling but somehow select for performance only those behaviours which are sex-appropriate.

The differential socialization explanation suggests that sex differences in behaviour occur as a result of the differential reinforcement of male and female children by parents and other significant adults. Maccoby and Jacklin outline four factors underlying the differential treatment of the sexes. Firstly parents may be reinforcing the child for behaviour that the parent considers to be ideal for a child of that sex. For example, if the parent feels that girls should be quiet and well-behaved, the female child would be rewarded for quiet, orderly behaviour and punished for loud, disorderly behaviour. The child may also be reinforced in accordance with what the parents consider to be behaviour characteristic of the child's sex. A parent thinks that girls are naturally quieter than boys. A girl, therefore, is rewarded for playing quietly and behaviour which departs

from this norm is more apt to be noticed and discouraged. Whether or not a child and parent are of the same sex may also influence the socialization process. Maccoby and Jacklin suggest that the adult expects to be a model for the same-sex child, tends to identify more strongly with him or her and therefore has higher expectations of him or her. The fourth factor influencing differential socialization stems from the child: Maccoby and Jacklin suggest that male and female children may stimulate their parents differently and so elicit different treatment from them. Boys, for example, may give cues that they are more active and aggressive than girls. Parents consequently respond by treating their male children differently, thereby encouraging the development of differential sex-typed behaviour. It is also possible, from this viewpoint, that sex-typed behaviour may emerge despite parental disapproval, e.g. aggressive behaviour.

Maccoby and Jacklin are inclined to downplay the influence of all the factors described above because they generally find little evidence for differential socialization, concluding that there is "a remarkable degree of uniformity in the socialization of the two sexes" (p. 348). They unexpectedly find, also, that boys undergo a more intense socialization experience than girls, receiving both more punishment (Smith and Daglish, 1977) and more positive feedback (Serbin, O'Leary, Kent and Tonick, 1973). To integrate these findings, Maccoby and Jacklin suggest that boys receive more attention in general than girls: "Adults respond as if they find boys more interesting, and more attention-provoking than girls." An activity level hypothesis provides a simple explanation for adults' more intense response to male children

in that "boys are more active, thus providing more stimulation to observers." Rather than consisting of overt attempts at shaping the child, Maccoby and Jacklin conceptualize the socialization process as a more subtle interaction between the attitudes and behaviours of the parent on the one hand, and different behavioural cues (e.g. activity level) of the male or female child which elicit differential amounts of parental attention on the other.

As simple and persuasive as the activity level hypothesis is, and in spite of a summary table of studies which generally find males to be more active than females (pp. 173-175), Maccoby and Jacklin go on to dismiss the activity level hypothesis as untenable on the grounds that no reliable sex difference in activity level has been proven. They reach this conclusion because sex differences in activity level have not been observed at all ages. For instance, few studies have concentrated on sex differences in adults.

Maccoby and Jacklin also discount the activity level hypothesis because those studies reporting significant sex differences examined activity level only in the context of situations which provided salient elicitors of activity for males and not for females. Consequently the sex differences discovered were situation specific. Two hypotheses are advanced to explain why this situational specificity results in finding a higher activity level in male subjects. The first hypothesis assumes that under stress females freeze and become less active, while males maintain their normal activity level. Presumably most of the studies were conducted in situations which were stressful to the subjects. A study by Maccoby and Feldman (1972), however, casts doubt on the stress

hypothesis in that children of both sexes tended to show reduced activity levels under stress. More recently, Eaton and Keats (1982) questioned the stress hypothesis when they found that both sexes showed reduced activity under the stress of being alone in a relatively novel setting. The second hypothesis, the peer hypothesis, assumes that differences in activity level are more frequently observed when children are in the presence of their peers than when they are alone. More specifically, Maccoby and Jacklin speculate that the presence of other males stimulates greater activity in males than the presence of peers does for females. Support for this hypothesis comes from a study by Pedersen and Bell (1970). Eaton and Keats (1982), however, failed to find a significant sex by condition interaction when activity level was observed in both an alone and same-sex triad condition; males were more active than females in both conditions. Thus the evidence for Maccoby and Jacklin's situation-specificity argument is sparse and far from convincing.

Child Effects and Adult Control

Although Maccoby and Jacklin dismiss the hypothesis that differences in activity level contribute to differential socialization of the sexes, some theoretical support for the activity level hypothesis comes from Bell and Harper (1977). Bell and Harper emphasize the role of the child's own behavioural cues in eliciting parental behaviour. In a chapter which re-examines research findings on socialization, Bell concludes that most of the evidence is in favor of a congenital contribution to hyperactivity. He uses hyperactivity as one example of

a causal influence which emanates from the child, pointing out that a parent tends to respond differently to the child who is overly active than to the child who is normally active. Bell describes a control theory model for parent-child interactions in which upper-limit and lower-limit control behaviours are elicited from parents in response to particular child behaviours. Control behaviours are a result of the selective activation of elements in the parent's repertoire of behaviours. Whether or not a particular control behaviour is activated is a function of the parent's whole past experience in a given sphere of interaction, including his values, stereotypes and past history of parent-child interactions. A particular parent-child interaction therefore includes the child's behaviour, which may have a significant congenital component, and the parent's control behaviour selectively activated from a whole repertoire of control behaviours. Control behaviours are activated when the child's behaviour violates certain limits. According to Bell, "each participant in a parent-child interaction has upper and lower limits relative to the intensity, frequency or situational appropriateness of behaviour shown by the other" (p. 65). A child behaviour which reaches the upper limit in terms of what the parent can tolerate results in the activation of upper limit control behaviour in which the parent acts to redirect or reduce the excessive behaviour. If a lower limit is reached the parent uses lower limit control behaviour to stimulate an increase in the insufficient behaviour.

Bell's control theory complements the activity level explanation for sex differences in that it provides an explanation for research

findings indicating that males generally receive more attention and undergo a more intense socialization than females. If males are more active, their higher activity level would be one factor likely to elicit more upper limit control behaviour from their caretakers. Bell describes upper limit control behaviour as "distraction, quick tangible reinforcement or nonreinforcement, holding, prohibiting verbalizations, and physical punishment." If girls are less active, their caretakers would be more likely to respond with lower limit control behaviour which includes: "drawing attention to stimuli, positively reinforcing increases in activity, urging, prompting, and demanding increased performance." The two types of control behaviour would result in a qualitatively different socialization experience for the two sexes.

Activity Level and Sex Differences

Given that activity level may be a potent child effect prompting differential adult responding, is there evidence for sex differences in activity level? If there are no sex differences, activity level as an explanation for sex-typed differential responding becomes questionable. Maccoby and Jacklin concluded that the evidence on activity level was ambiguous. However, methodological flaws in the collection and integration of research data tend to weaken their conclusion that there are no reliable sex differences in activity level.

In order to summarize the data on activity level Maccoby and Jacklin used a box-score approach to compile a table of research findings indicating whether between-sex comparisons of activity level produced a significant or non-significant result favoring one sex over

the other. Block (1976) evaluated their technique by examining the proportion of significant findings relative to the total number of comparisons included in the table. Of the 50 comparisons tabled, Block reported that 30% found males to be more active and 6% found females to be more active. To strengthen the case for sex differences in activity level, Block included for analysis an additional 9 studies containing between-sex comparisons of activity level (taken from the annotated bibliography of The Psychology of Sex Differences) which Maccoby and Jacklin had failed to include in the activity level summary table. With the addition of the 9 studies cited by Block, the percentages change to 42% of studies reporting significant differences with males being more active, and 5% of studies reporting females as significantly more active, percentages which are comparable to percentages for sex differences considered well-established by Maccoby and Jacklin.

Block also criticized the box-score approach used by Maccoby and Jacklin because it failed to take the variability in power of individual comparisons into account. Studies with large and small sample sizes are grouped together in the final analysis in spite of the fact that it is easier to detect a significant difference with a large sample size than with a small one. Block also points out that the box score approach makes no provision for reliability of dependent measures. The Maccoby and Jacklin summary table on activity level includes studies which employ a variety of measures from actometers to rating scales. Each study has an equal weight in the integration whether or not satisfactory reliability data for the measure is reported. There are now better alternatives for research integration and these newer procedures are outlined and discussed below.

Research Integration Techniques

The traditional method of research integration has been the narrative review in which relevant empirical studies are collected and "poorly designed" studies discarded. Conclusions from the remaining studies are compared and those with consistent results retained. Findings or themes which appear frequently are included in the final integration while those findings which appear only once or twice may be excluded whether or not they come from studies which are comparable in design. The narrative review method has been criticized on the grounds that much valuable information is likely to be discarded. A source of bias is introduced in that the organization and integration of the relevant studies frequently has to occur before the actual message from the studies can be extracted. Discarding the flawed studies may result in "those remaining being one's own work or that of one's students or friends" (Glass, 1976).

A better, more systematic approach is the voting method which is similar to the box score method used by Maccoby and Jacklin. All the studies which have data on an independent variable of interest are examined. Possible outcomes--significant or non-significant--are defined and the number of studies whose results fall into each category are counted. The modal category is then considered to give the best picture of the relationship between dependent and independent variables. As noted earlier, the voting method has been criticized on a number of counts. Unlike the narrative method, design characteristics of the studies are not considered in the integration. In addition, use of the voting method may produce a final summary which is biassed in the

direction of those studies which employ a large sample size. This can occur because studies with larger sample sizes are more powerful and more likely to produce significant results. With the voting method, sample size is not taken into consideration and studies with large and small sample sizes alike have one vote. Perhaps more importantly, the voting method makes no allowance for the size and strength of the relationship between independent and dependent variable. A strong relationship has the same effect as a weak one in tallying up the votes; each study has only one vote depending on whether or not its results are significant.

Glass (1976) recommends the use of another technique, meta-analysis, as a statistical method for the integration of research results. Meta-analysis is simply defined as "the integration of research through statistical analyses of the analyses of individual studies" (Smith and Glass, 1977, p. 752). Glass's own method is to calculate an effect size for each study by taking the mean difference between treated and control subjects and dividing by the standard deviation of the control group. An average effect size and standard deviation of the effect size can be calculated across all studies. The Glass technique effectively handles major criticisms of the more traditional techniques described earlier. It reduces the amount of sujective bias which can be introduced into the integration by using standard statistical procedures. Meta-analysis also takes into account the issue of strength of relationship between independent and dependent variables by assessing the size of the effect being studied (Cooper, 1979). Finally, the power and sample sizes of the individual studies

which enter into the integration directly influence the effect size statistic which is calculated by using a standard deviation.

Meta-analysis seems more inclusive than the alternatives, but it too has been criticized on a number of different grounds. Gallo (1978) points out that in a meta-analysis many different kinds of dependent measures are aggregated together; in the Smith and Glass case, for instance, "ranging from elaborate clinical judgments to scores on pencil-and-paper tests." It is difficult to extricate any meaningful information from "such a hodgepodge," and even if it were possible to extricate it, it is difficult to interpret it. Does a score of .68 of a standard deviation above a control group on a measure with poor reliability mean anything? Can it be meaningfully compared with another score on a different dependent measure whose reliability is well-established?

Eysenck (1978) criticizes meta-analysis on the grounds that it does not take into account the quality of the design of the studies which go into it. Poorly designed and well designed studies are all given equal weight in the final integration. Eysenck harshly criticizes the Smith and Glass meta-analysis on psychotherapy outcome studies for being "a compilation of studies mostly of poor design, relying on subjective, unvalidated, and certainly unreliable clinical judgments, and dissimilar with respect to nearly all the vital parameters."

An additional problem is that of selection bias which may occur if a large number of studies are excluded from the meta-analysis in a systematic way. For example, if it is true that significant results are more likely to be accepted for publication than non-significant results,

a meta-analysis done on the basis of the published literature may be biassed in the direction of significance.

A fourth problem which occurs is the problem of non-independent results. Many studies employ more than one dependent measure and so involve more than one test of the relationship between independent and dependent variables. Such a study will yield more than one finding which can be used for meta-analysis, and the data then violates the independence assumption. On the other hand, excluding some of the findings or averaging across the study results in the loss of useful information.

Finally, there is the more global criticism that all aggregations of research are misleading. Gallo (1978) points out that an effect size is meaningless unless there is a context by which to evaluate it--"the same size effect can be incredibly important, or almost totally unimportant, depending upon the context." In addition, in aggregating results, there is a tendency to ignore all but the main effect of the independent-dependent variable relationship under investigation. Important interaction effects or relationships with variables other than the one specified are ignored. Light and Smith (1971) maintain that unless research is integrated at the raw data or primary analysis level, important systematic patterns in the data (such as large differences in variance between treatment and control group when only means are reported) may be missed at the secondary analysis level.

In response to the criticism that in a meta-analysis "apples are being mixed with oranges" and results of studies with different outcome measures and experimental designs are being mixed together, Smith and

Glass (1977) reply that integrating studies with different outcome measures is defensible. In their meta-analysis on psychotherapy outcomes, they give three reasons why they feel this to be the case. In the first place, all outcome measures are related to the same construct, in this case, psychological well-being. Secondly, it is necessary to mix different outcome measures for the sake of practicality; it is important to get an answer to a specific question. Glass (1977) points out that there is no need to integrate or compare studies which are the same; it is different studies that need to be integrated. Finally, each independent researcher has already made a value judgment concerning which dependent measure or definition of the construct was best suited for his particular study. Smith and Glass see no reason to repeat the process at second hand. It should also be noted that if aggregations are inherently misleading, psychologists should stop calculating means from the responses of different individuals.

In response to Eysenck's criticism that a meta-analysis does not take into account quality of experimental design, Glass and Smith (1978) respond that errors in design are not so critical when dealing with a large number of studies since measurement errors tend to average out in groups, provided no systematic bias is at work. If a systematic bias is suspected, then regression analysis can be used to control for it, such as the subjectivity variable entered into the multiple regression analysis done by Glass and Smith. Furthermore, Glass states that there is too much valuable information in so-called poorly designed studies for them to be discarded and he wonders "whether well-designed and poorly-designed experiments give very different findings" (Glass, 1976, p. 4).

The danger of systematic selection bias makes the use of statistical procedures, such as random sampling, essential. Ideally, studies for the meta-analysis should include the universe of all studies in existence, published or unpublished or a random sample of this universe. If access to many of the studies is not possible, and this typically is the case for unpublished research, Rosenthal (1979, 1980) describes how to calculate a fail-safe N, an estimate of the number of additional studies showing a null relationship which are needed to increase the probability of the results of a meta-analysis to above the .05 level of significance. If the existence of just a few studies showing null results is all that is needed to raise the level of probability to greater than .05, it is reasonable to conclude that the results of the meta-analysis are not very robust, and that selection bias may constitute a serious problem. If many studies are required it is unlikely that selection bias can seriously affect the results.

Glass (1977, p. 375) points out that there is no simple answer to the problem of non-independent results. The researcher must decide how many different independent units of information exist in the data. The simple solution, according to Glass, is to regard each findings as independent, whether or not there are several findings per study, and to bear this in mind when interpreting the results of the meta-analysis. A more complex procedure is to use some method of averaging the findings in a single study. This achieves independence of the data at the expense of loss of information: it becomes impossible to assess the relationship between magnitude of effect and type of measure. It would also be possible to combine approaches by using only independent effect

sizes for estimating the basic relationship between independent and dependent variables while retaining all effect sizes for determining the relationships between magnitude of effect and other study characteristics.

Glass's response to the point that, by definition, integrations of research have to be misleading is that, while it is true that some information is lost in the process, practicality is also important. In a field which is inundated with pieces of diverse information, it is critical to "find the knowledge in the information" (Glass, 1976, p. 4). One of the major advantages of meta-analysis is that it is possible to retain and make use of more information than other integration techniques. The information doesn't need to be interpreted as a finished solution to a research problem, but the findings can be used as guidelines for future research. Also, correlational analyses performed on the whole data set or on various subsets can be used to investigate the specific independent and dependent variables of most interest to the researcher. In the meta-analysis on psychotherapy outcome research, Smith and Glass (1978) employed regression analysis to estimate the interaction effect of therapist experience by type of client (diagnostic category). Light and Smith's (1971) case that raw data re-analysis represents the soundest data integration technique is valid. Raw data re-analysis, however, is time-consuming and impractical, if not impossible in most cases, because of the inaccessibility of the raw data. Much valuable information from many older studies whose raw data has not been retained would have to be discarded.

A meta-analysis applying the "quantitative rigor" of statistical analysis to the integration of research data (Glass, 1976) seems the appropriate methodological choice to replace Maccoby and Jacklin's box-score method for investigating the relationship between sex and activity level for the following reasons:

1) The selection of studies for inclusion into a meta-analysis tends to be less systematically biassed than selection through the traditional narrative and box-score approaches. Ideally meta-analysis is performed on a random sample of available studies, and a fail-safe N (Rosenthal, 1979, 1980) calculated to assess the possible effects of unavailable studies.

2) Meta-analysis takes into account the number of subjects and therefore the power of the individual studies that enter into the integration. The box-score approach makes no attempt to allow for differences in power of the individual studies and the traditional narrative approach does so in an idiosyncratic way that is open to researcher bias.

3) In using the narrative approach, the reviewer typically has to exclude many studies in the interests of manageability. It is possible to be much more inclusive with meta-analysis and final results are based on the individual findings of many studies.

4) Use of the effect size statistic employed in meta-analysis makes it possible to estimate the magnitude of the

relationship between activity level and sex. Neither the narrative review method nor the box-score approach address this issue of relationship strength.

5) Using correlational procedures it is also possible to investigate and estimate the size of the relationship between effect size magnitude and relevant study characteristics, allowing for exploration of factors which influence the size of sex differences in activity level. The study characteristics are coded and treated as independent variables in the correlational analysis.

Study Characteristics Influencing Effect Size

According to the literature, age is one of the substantive characteristics whose relationship to activity level should be investigated. According to Maccoby and Jacklin, sex differences in activity level vary in size and direction according to age, leading them to hypothesize that sex differences in activity level may be age-specific. If this hypothesis were true, it was expected that results of the meta-analysis would show that larger effect sizes were found with samples of preschool age and smaller effect sizes were found with younger and older samples. Results of the meta-analysis were also expected to lead to suggestions for direction of further research if, as expected, there was a comparative lack of activity level studies which used adults or very young infants as participants.

Some of the more important questions about sex differences in activity level focus upon characteristics of the settings of the

studies. Unpublished data from Eaton, Nottelmann, Williams and Williams (Note 1) suggest that there is a relationship between restrictiveness of the setting and activity level in children. Observational data was collected in two elementary school classrooms, one less structured and characterized by more free-choice time, the other, more formally structured and characterized by more teacher-directed time. Differences between the sexes in play behaviour (which was characterized by gross motor movement, toy involvement, and fantasy activity) differed from classroom to classroom, with the sex difference in play behaviour being greater in the unstructured classroom (boys>girls). The coding category "restrictiveness of setting" was designed to collect information to deal with this issue. The "type of setting" category was also designed to give this kind of information (i.e. a structured lab setting would be more restricted than a home setting). In addition, it was expected that categorization of the type of setting employed would give information which could be used to determine whether the findings of studies conducted under naturalistic conditions (home or school) were correlated with larger differences between the sexes than were studies conducted in the lab.

Maccoby and Jacklin have hypothesized that stressfulness of setting is one variable that may maximize sex differences in activity level, the theory being that females tend to freeze more than males when under stress. They dismiss the hypothesis on the basis of findings to the contrary, but few studies (Maccoby & Feldman, 1972; Eaton & Keats, 1982) address this issue directly. It seemed important to gain more information on this question; the "stressfulness of setting" variable, a

coder-judged rating scale of overall setting stressfulness, was therefore designed to assess stressfulness. Similarly, a "novelty of setting" category was used to collect information on whether a study's setting was unfamiliar and, consequently, stress-inducing for the subjects.

Maccoby and Jacklin also hypothesized that peer presence may be a factor in attenuating sex differences with males being stimulated to a higher level of activity than females by the presence of same-sex peers. Evidence on this point is contradictory; Pedersen and Bell (1970) supporting it and Eaton and Keats (1982) failing to find greater sex differences in the presence of same-sex peers. In order to collect further information on this issue, it was decided to code studies for peer presence.

Various relevant methodological characteristics such as obtrusiveness of measure, reliability coefficient, number of raters, etc. were also coded. Information collected on these variables acted as a quality control check on studies included in the meta-analysis, helping to circumvent the criticism that meta-analyses may include studies of seriously flawed design. If the design quality of the study has a strong impact on the magnitude of the obtained effect size, smaller effect sizes would be correlated with less reliable measures. This information could, in turn, suggest methodological changes for future research.

From a preliminary survey of the literature, type of measure emerges as a crucial issue. Maccoby and Jacklin observe that different kinds of measures are used to estimate activity level at different ages

so that it is difficult to get reliable information on sex differences. Reliability, Epstein (1979) argues, is very important because it is only possible to detect the existence of stable individual differences by averaging over a sufficient number of occurrences of behaviour, something that most studies have neglected to do. With this in mind, the length of the behavioural sample employed in each study was coded.

Activity level was assumed to be a dimension of temperament which cuts across specific behaviours. Inclusiveness of activity level measure referred to whether or not the measure used in a particular subject was based on a broad and inclusive sample of the subject's behaviour (e.g. a global teacher rating of a child's activity level) or on a narrow, specific sample (e.g. number of times a child crosses into an adjacent grid on a marked floor). The literature suggests that the more inclusive measure would be the more powerful. For observational data, the longer the behavioural sample (coded here in minutes), the greater the likelihood of finding significant sex differences. For rating scales, similarly, the greater the number of scale points per item and the more raters, the more reliable and powerful the scale. Studies which report reliability data in the form of reliability coefficients were also expected to be associated with maximal sex differences in activity level.

Information on type of measure (e.g. rater vs. mechanical) used in the calculation of the individual measures was used to determine if sex differences were found with some types of measures but not others. If this were the case, one would expect larger sex differences for rating scales than for more objective measures since Maccoby and Jacklin have

suggested that rater bias in the form of sex-role stereotyping may contribute to findings of large sex differences. A complication arises, however, because raters are typically making a judgment on the basis of a great deal of behavioural data. If the inclusiveness hypothesis were true, effect sizes may be larger for ratings than for objective measures because the ratings are more inclusive. On the other hand, if activity level is a general disposition with a robust influence, sex differences should emerge regardless of the method employed.

Investigator's description of sample selection was coded to provide information on the existence of selection biases in the data and to determine whether additional studies focussing on a greater variety of subjects was needed.

One code, percentage of male authorship, was added to the coding scheme after publication of Eagly and Carli's (1981) report of a significant relationship between the percentage of male authors for a publication and the size of observed sex differences in persuasibility. If male authors are unconsciously motivated to perpetuate sex differences, larger sex differences should be associated with a high percentage of male authorship.

Summary

On examination of the research evidence, the issue of sex differences in activity level remains unresolved. Resolution of this question is important because a child's activity level could constitute a powerful "child effect" which in turn may influence adult behaviour directed towards the child. If sex differences in activity level exist,

such differences might, in part, explain why boys and girls undergo differential socialization, thereby learning to behave according to conventional sex-role norms. The purpose of the present study was threefold: 1) to collect as many studies testing the difference in activity levels between males and females as possible; 2) to determine from this research sample whether sex differences in activity level exist and to estimate their size and direction; 3) to determine which study characteristics have the strongest association with magnitude of sex differences in activity level.

Method

Study Selection Procedures

An empirical study was selected for inclusion in the meta-analysis if one or more of its dependent measures was a measure of gross motor activity level or a closely related construct, if effect sizes for sex differences could be calculated from the available information, and if there was no indication that subjects were characterized as non-normal.

A dependent measure was defined as any measure of gross motor activity level the researcher chose to employ including mechanical, observational and rating scale measures. Studies which investigated energy level, motility, vigor of play, rhythmicity, etc. were all considered for the meta-analysis if, by the primary researcher's definition, those variables were closely related to gross motor activity level.

Findings which were based on a non-normal sample were excluded from the meta-analysis. A non-normal sample was defined as any sample which

consisted of subjects who had been diagnosed as having emotional or behavioural problems or who were physically or mentally disabled. Studies which used samples of hyperactive children, psychotics, the mentally retarded or samples drawn from any other group with characteristics possibly affecting activity level were therefore excluded. In some studies focussing on non-normal subjects, normal control groups were employed, and data from these control groups was included when possible.

Studies which did not contain the information necessary to derive an effect size were not included in the meta-analysis. It was decided not to write away for missing data because there is evidence that requesting original data is both time-consuming and unproductive. Glass (1977) reports that one researcher requested original data from 37 authors who published in 1959-1961. Of these, 5 did not reply, 21 reported data lost or destroyed, 2 claimed proprietary rights and 9 sent data (4 too late to be useful).

Computer search procedures were used to locate relevant studies. A free text search was conducted on the Psychological Abstracts data base.[1] Additional studies were located through the standard procedure of following up references from relevant review articles and empirical studies. Some additional studies were found solely on the basis of chance, for example, through colleagues who knew about the research and who, in the course of their own work, noticed studies on sex differences in activity level.

---

[1]The search was conducted in June, 1980 using the following descriptors: rhythmicity, movement, activity level, motility, restlessness, vigor in play, motoric activity, motor activity, tempo, locomotor activity, energy level, change in activity.

All search procedures were restricted to the published English language literature. This publication restriction was made because of the difficulties inherent in translation and in obtaining unpublished work. Despite this restriction, selection bias should still have been minimal because, although studies finding significant results are more likely to be published, the results are equally apt to be significant in either direction (male>female or female>male). Consequently, if no real significant differences exist, positive and negative significant findings should cancel out. Furthermore, analyses for sex differences in activity level are often peripheral to the study's central purpose and publication decisions are probably not based on activity level findings. Bias on the part of the primary researcher to include only significant results with a specific directional effect should therefore have been minimal as the analysis for sex differences was usually secondary in interest to the main hypothesis. Finally, it was expected that restricting the sample to published studies would act as a control for quality, since better designed studies are more likely to be published.[2] As a precautionary measure, a fail-safe N was calculated according to procedures outlined in Rosenthal (1979, 1980) to give an estimate of the number of unincluded and/or unpublished studies showing null results needed to offset the findings of the meta-analysis.

According to the above criteria, 41 studies based on data from 31,698 subjects were located and included in the meta-analysis. This represented a ratio of 1 study included for every 5 studies initially screened. The majority of located studies simply did not have the

---

[2]Wide variability in the quality of published work was, however, apparent in the reading of studies.

information needed or did not involve an assessment of gross motor activity level.

Coding

The selected studies were coded for the presence of certain substantive and methodological features suggested by the literature as interesting. Substantive or theoretically interesting variables included age, peer presence, stressfulness of setting, novelty of setting, restrictiveness of setting,etc. Methodological features of interest included reliability, number of data points per rating scale (number of scale points per item and number of items on scale), sample size, etc.[3]

To facilitate the assessment of measurement reliability for this coding system, two coders coded the entire sample of studies for all predictor variables according to the conventions and definitions outlined in Appendix B. Coders independently rated a small set of studies, assessed agreements and disagreements and revised definitions where necessary to improve reliability and to handle cases which were not satisfactorily dealt with by the coding manual. The process was repeated for the entire set of 41 studies and allowed for the calculation of reliability for all of the codes, except the "percentage of male authorship" code. This variable was added to the coding scheme after all other coding had been completed, and only one rater coded this category.

---

[3]See Appendix B for definitions and conventions used in the coding of study features, and see Appendix C for a sample of the coding form.

An analysis of the coded study characteristics led to the following description of the typical study included in the meta-analysis. The modal study selected for inclusion was published in a journal in 1970-1972. Typically a gender difference in activity level was not a major focus of the study. Fifty per cent of the authors of the study were male. Findings from the modal study were based on a sample of 69 subjects 5 to 6 years old who were tested in a preschool setting with at least one adult and a number of peers present (it was difficult to determine whether the peers were same- or mixed-sex). The setting was judged to be familiar to the subjects and low in both stressfulness and motor restrictiveness. The typical measure was a highly inclusive and unobtrusive rating scale and raters tended to be either parents or teachers.

Results

Effect size statistics were calculated for each of the studies in the meta-analysis using the general formula given by Cohen (1969): d=(Mean of males - Mean of females)/Standard deviation. In other words, an effect size was defined as the difference between the means for male and female subjects divided by a pooled standard deviation.

The effect size formula requires means and standard deviations for the male and female groups or a pooled standard deviation for both. In cases in which means and standard deviations were not reported, effect sizes were calculated (Glass, 1980; Smith, Glass & Miller, 1980) from *t* and *F* ratios together with sample sizes for each sex. Available

formulas were also used to calculate effect sizes from data reported as correlations or proportions.

The total sample of 41 studies yielded 104 individual sex difference comparisons expressable as effect sizes. Comparison rather than study was chosen as the unit of analysis on the grounds that the danger of losing information outweighed the danger of introducing dependency into the data. To counter some of the problems raised by non-independent effect sizes in cases where multiple measures were used on the same sample, or where the same measure was used more than once on the same sample at different ages, a mean was calculated from the non-independent effect sizes. The 41 studies thus yielded 54 independent effect sizes.

Both non-independent and independent effect sizes were used to investigate the following hypotheses suggested by the activity level literature:

1) Average effect size would support the hypothesis that males have a higher activity level than females.
2) According to Maccoby and Jacklin, sex differences in activity level are age-specific. If this were so, it was expected that larger effect sizes would be correlated with pre-school samples while smaller effect sizes would be correlated with younger and older samples.
3) Restrictive settings would reduce activity level for both sexes leading to smaller effect sizes.
4) Sex differences would be more likely to be detected in stressful situations; therefore effect size would be

correlated with stressfulness. Also, studies conducted in settings which were novel to the subjects and therefore more stress-inducing would be associated with larger effect sizes because activity level in females would be suppressed.

5) Presence of peers would have a weak relationship with effect size. This is consistent with Eaton and Keats's (1982) finding of no sex by condition interaction when children were observed at free play during alone and peer presence conditions.

6) Reliability of measurement was expected to correlate positively with effect size. Studies which employed measures which were more inclusive (those which examine a broader spectrum of behaviour) would produce larger effect sizes. For rating scales, larger effect sizes would occur in conjunction with a greater number of items on the scale, a greater number of scale points per item and a larger number of raters. For observational data, larger effect sizes would be positively associated with size of behavioural sample.

7) It was expected that larger effect sizes would be found in studies which employed rating scales as the dependent measure. This outcome would be consistent with Maccoby and Jacklin's suggestion that rater bias in the form of sex role stereotyping contributes to the reported sex differences as well as with the hypothesis that more inclusive measures would uncover larger effect sizes.

Are Males More Motorically Active Than Females?

In the first phase of data analysis, the calculation of summary statistics, mean and median effect sizes based on the 54 independent effect sizes were calculated. As expected, results were in the direction of males having a higher activity level than females, with the male mean being .52 standard deviations larger than the female mean. Another way of stating this difference is to compare the overlap of the male and female distributions. The average male (at the 50th percentile) is more active than 69% of females (from tables given in Cohen, 1969). In terms of variance, 5.9% of the variance in gross motor activity level is attributable to sex differences (Cohen, 1969). Because the distribution of effect sizes was positively skewed, medians were also calculated. Replacing the mean with the median, the male median was .40 standard deviation units higher than the female median indicating that the average male (50th percentile) is more active than 66% of females. Using the median, the amount of variance in activity level attributable to gender differences is 3.8%.

These data were tested for significance in two ways: through the construction of confidence intervals around mean and median effect size and through the use of estimated z-scores calculated from the effect sizes. Using the standard error of the mean (.06), 95% confidence intervals ranged from .40 to .64 indicating a non-null difference at $p<.05$. Using the median as a measure of central tendency, confidence intervals ranged from .28 to .52, again indicating a non-null effect.

An additional test was done using a second meta-analytic technique, Rosenthal's (1978) method of combining probabilities. Estimated z-

scores were calculated for each of the 54 independent effect sizes (see Table I) according to a formula given by Rosenthal (1979) and summed. The combined probability expressed by the z-scores (19.75) was highly significant ($p<.0000001$). Excluding the effect size obtained from a study by Stone (1981) which had a very large sample size (N=25,000), the combined z was 15.97, still very significant.

Could Undiscovered Studies Change the Conclusion?

Estimated z-scores used for the significance tests were also used to calculate a fail-safe N (Rosenthal, 1979). As stated previously, the fail-safe N gives an estimate of the number of unincluded studies showing null results (effect size=0) necessary to offset the findings of a meta-analysis. It was used to estimate the effects of bias introduced into the activity level meta-analysis by the decision to restrict the sample only to published studies thereby testing the robustness of the finding that males have a higher activity level than females. It was found that 7723 unpublished or otherwise unincluded comparisons (independent effect sizes) would be necessary to reduce a combined z of 19.75 to less than a .05 level of significance. Excluding Stone (1981) the file drawer N became 3650. Considering the difficulty encountered in locating the 41 studies included in the current sample, it seemed clear that the results of the activity level meta-analysis were not susceptible to the file drawer problem.

Reliability of Coding Scheme

Before doing a correlational analysis of the data, reliability of coding was assessed through the calculation of inter-rater correlation coefficients for continuous variables and Cohen's (1960) kappas for categorical variables. Since rater judgments were combined prior to any analyses, the reliability of the combined codes was then estimated through calculation of Spearman-Brown coefficients (See Tables II and III for inter-rater, kappa and Spearman-Brown correlation coefficients for all coded predictor variables.) The reliability of the combined data was quite high, ranging from .64 to 1.00 (Spearman-Brown corrected), the only exception being number of items on scale (.23).

Correlational Analyses

The next phase involved analysis of correlations between the coded predictor variables and the effect sizes. For these analyses, all 104 effect sizes were used because coding was done for each individual comparison, rather than for each study. Before beginning this phase, the distribution of effect sizes, which was positively skewed, was normalized using log transformations. Simple correlation coefficients were then calculated between all predictor variables on the one hand and normalized effect sizes on the other (See Tables IV and V for a listing of correlation coefficients, associated probabilities and sample size for all predictor variables.)

A number of the methodological predictors proved to be significantly correlated with effect size. Basis for selection ($r=-.20$, $p=.05$) was negatively correlated with effect size, indicating that

subjects chosen from public schools or nursery schools and day care centres showed larger sex differences in activity level. A possible reason for the significant correlation is suggested by the inter-correlation between the basis for selection and log of mean chronological age predictor variables ($r=-.19$, $p=.06$). Large effect sizes tended to be associated with younger subjects who were, of course, recruited from schools and day care centres more frequently than from community organizations and other sources.

Type of measure ($r=.16$, $p=.12$) showed a tendency towards a positive correlation with effect size initially suggesting that kind of measure used (rating scale mean=.43, quantitative mean=.44 or observational mean=.71) affected the size of the obtained sex difference. The size of this correlation, however, seems to have been affected by the skewness of the distribution of effect sizes. An examination of the medians of the three groups shows that type of measure is not associated with sex differences in activity level (rating median=.38; quantitative median=.40; observational median=.41).

Consistent with the hypothesis that males have a higher activity level than females, direction of effect ($r=-.32$, $p<.001$) was negatively correlated with effect size. That is, large effect sizes were correlated with the finding that male activity level is greater than female activity level and smaller effect sizes were correlated with the reverse finding.

Also as expected, significance of results as reported by primary author was correlated with sex differences in activity level ($r=-.55$, $p=.0001$). Effect sizes based on results which were reported as

significant by the authors were larger (median=.59) than were effect sizes based on non-significant results (median=.18). Similarly, probability level as reported by primary author was negatively correlated (r=-.47, p=.003) with effect size indicating that high probability of outcome was associated with small effect size and low probability was associated with large effect size.

A final methodological predictor that correlated with effect size was the kind of statistic used to calculate effect size (r=-.29, p=.003). Larger effect sizes resulted from calculations from t-statistics (mean=.75) and smaller ones were obtained from means and standard deviations (mean=.36). This finding might have reflected the influence of hypothesis testing, since smaller effect sizes were obtained directly from the raw data (means and standard deviations). That most effect sizes (58 of 104) were calculated from the raw data supports the overall robustness of the results of the meta-analysis.

Contrary to expectation, other methodological predictor variables proved to be unrelated to effect size. Among these were the variables associated with reliability of measure (number of scale points per item and items per scale, number and description of raters, number of observers, length of behavioural sample in minutes, type and size of reliability coefficient) and the percent male authorship variable.

Of the substantive predictors, only three proved to have significant correlations with sex differences in activity level. Minimum and maximum age of subjects were both positively correlated with magnitude of effect size (r=.29, p=.03; r=.23, p=.03) although the correlation between mean chronological age and effect size barely missed

significance (r=.18, p=.06). The results point to the interpretation that, as subjects increased in age, sex differences in activity level increased. These findings have to be interpreted in view of the fact that the distribution of subject ages was badly skewed and few studies used adult subjects.

The correlation between number of peers present and effect size showed a tendency towards significance (r=.42, p=.17), suggesting that the magnitude of sex differences in activity level increases as number of peers increases. The correlation is based on only twelve cases, however, and other similar predictors (presence or absence of peers, and whether peers are same- or mixed-sex) are not significant. It is difficult, therefore, to interpret the findings as support for the same-sex peer presence hypothesis of greater activity level in males. The suggestion that peer presence (whether same- or mixed-sex) may differentially affect the activity level of the two sexes seems to be worth further investigation.

Another substantive predictor, restrictiveness of setting, showed a tendency towards significance (r=-.17, p=.11). Although not significant, the correlation was in the expected direction: smaller sex differences in activity level were obtained in more restrictive settings.

The lack of correlation between the "test central to hypothesis" predictor and effect size (r=-.03; p=.76) suggests that researchers' bias towards finding a significant sex difference in activity level did not affect the results of the meta-analysis. A test of sex differences in activity level was a major hypothesis of the study in only 25 of 104

cases and a peripheral hypothesis in over 79 of the 104 cases. Both types of studies were, however, associated with similar effect sizes (mean ES of yes group=.52, median=.32; mean ES of no group=.48 median=.40). Researchers whose central purpose was presumably to find a sex difference in activity level were no more apt to find large effect sizes than were their colleagues who reported activity level data incidentally.

Other substantive variables hypothesized to be important predictors of sex differences in activity level (stress of setting, novelty of setting and presence of peers) proved to have non-significant correlations with effect size.

Discussion

The major hypothesis that males have higher gross motor activity level than females is supported by the results of the meta-analysis. The sex difference is .52 of a standard deviation using the mean effect size or .40 of a standard deviation using the median effect size. Cohen (1969) suggests an effect size of .50 is moderate in size. Compared to other sex difference meta-analyses, the average effect size for activity level is moderately large and compares favorably with results from meta-analyses which focus on supposedly well-established sex differences. (See Table VI for a comparison of results with other sex difference meta-analyses.) For example, a meta-analysis by Hyde (1981) finds that females score higher by .35 standard deviations than males on measures of verbal ability. The findings of the Hyde meta-analysis are based on 12 comparisons, many fewer than the 54 independent comparisons on which the activity level meta-analysis is based.

Other data generated by the meta-analysis suggests that the finding of a sex difference in activity level is quite robust. For example, results of the correlational analysis between effect size and the "test" variable (whether or not the sex differences test constitutes the main hypothesis of a study) indicates that this probably did not effectively bias the overall results of the meta-analysis.

File drawer calculations also lend support to the hypothesis of sex differences in activity level. The fail-safe N of 7723 (3650 if Stone, 1981, with an N of 25,000 subjects is excluded) indicates that a great many comparisons averaging null effects are needed to overturn the results of the activity level meta-analysis. A thorough literature search resulted in finding only 41 usable studies (104 comparisons and 54 independent comparisons) so it seems highly unlikely that 7723 (or 3650) unpublished and/or unincluded comparisons testing sex differences in activity level exist.

Additional support for the robustness of the results comes from the finding that effect size seems to have little or no relationship with reliability and type of measure. Obtrusiveness (reactivity) of measure and inclusiveness of measure are also uncorrelated with effect size. These results suggest that supposedly more subjective, less rigorous measures such as global rating scales produce the same effect sizes as more rigorous and quantitative mechanical measures. Unpublished data by Eaton (in press) finds a similar correlation between mechanical and global rating scale measures of activity level. The correspondence in results across types of measures also suggests that observers/raters are not unduly biassed by social sex-role stereotypes. The data indicates

that any biases held by researchers on the basis of their own sex also have little to do with whether or not they find males to be more active than females.

The evidence relating to Maccoby and Jacklin's suggestion that sex differences in activity level may be age-specific is somewhat less clear-cut. The correlation between mean chronological age and effect size tends toward significance and the correlations between minimum chronological age and maximum chronological age and effect size are clearly significant. Interpretation of these findings must be tempered by the fact that the distribution of subject ages is badly skewed, with few activity level studies using subjects older than preschool age. With so few studies using older children and adults as subjects, it would be impossible to state whether or not the trend for increasing age to be correlated with larger effect sizes continues past primary school age. Interpretation of results is more difficult because, as Maccoby and Jacklin (1974) suggest, measures of activity level in children may be inadequate measures of activity level in adults. Gross motor activity expressed as physical movement (number of movements of arms and legs as measured by actometers) by a child may be expressed as more goal-directed mental and physical activity in the adult. More evidence on activity level in older samples would be more useful than additional studies with children.

The evidence in favor of the hypothesis that restrictive settings attenuate sex differences in activity level is also somewhat tentative. As the actual correlation is not significant, it is impossible to confirm or disconfirm the restrictiveness hypothesis. The negative

direction of a correlation which tends towards significance and an examination of mean and median effect sizes of the low, medium and high restrictiveness groups, however, shows a consistent pattern of maximal sex differences in less restricted settings and minimal sex differences in highly restricted settings. The evidence tends to support the restrictiveness hypothesis.

The hypothesis that presence of peers has little or no relationship to sex differences in activity level is supported by the data from the meta-analysis. Absence or presence of peers has no correlation with effect size. Similarly, whether or not peers are same- or different-sex does not seem to affect sex differences in activity level. The correlation between total number of peers and effect size, although significant, is based on comparatively few cases. The evidence tends to refute Maccoby and Jacklin's suggestion that male children are more active than female children in the presence of same-sex peers, but the results are ambiguous enough to warrant further investigation.

Results of the meta-analysis also show that stressfulness of setting does not affect sex differences in activity level, as effect size is not significantly correlated with coder judgements of stressfulness. Examination of mean and median effect sizes across low, medium, and high stress settings show sex differences of consistent size. These findings must be interpreted with caution, however, in view of the small number of included studies which measure activity level in highly stressful circumstances. Similarly, an examination of the data on novelty fails to corroborate Maccoby and Jacklin's hypothesis that sex differences in activity level may be affected by stressfulness

(including novelty) of setting. Again, results must be interpreted cautiously due to the lack of effect sizes obtained in settings which were truly novel to the subjects. On the other hand, most real life settings are not extremely stressful or novel, so stressfulness and novelty of setting are unlikely explanations for the existence of sex differences in activity level in the first place.

In summary, the findings do not support the argument that sex differences are specific to stressful or novel situations in which same-sex peers are present. The evidence for age patterns in activity level sex differences also remains ambiguous because of the relatively restricted age range in the surveyed studies. Instead, the data point clearly to the existence of a sex difference of moderate size in gross motor activity level. Although somewhat inconclusive, the tendency for large effect sizes to be associated with unrestricted, more naturalistic settings also argues for the existence of real sex differences that may be masked by more structured and restrictive settings. These findings are robust in that they are unrelated to measurement issues or to researcher bias as measured by researcher gender or expectation. A large fail-safe N renders implausible the contention that unincluded studies would overturn the results.

The pattern of results has clear implications for future research:

a) Studies using older school-aged children and adult subjects are needed to determine if age patterns in activity level persist into adulthood.

b) Studies which measure activity level across a wider variety of settings are needed to give further information

on setting specificity of sex differences. The modal study in the meta-analysis measures activity in school- or preschool-aged subjects in a school or preschool setting.

c) Research should be directed towards the study of sex differences in activity level in medium to highly stressful settings to further evaluate the hypothesis that sex differences are maximized under stressful circumstances.

d) Similarly, more activity level research should be conducted in settings which are less familiar and more novel to subjects.

e) The results of the meta-analysis suggest there is a tendency for males to be more active in the presence of peers, but few studies specify whether the peers are same-sex or mixed-sex. Also few studies directly compare activity level when measured in an alone condition with activity level measured in a group condition. In order to evaluate Maccoby and Jacklin's same-sex peer presence hypothesis, more research should address these issues.

In the process of obtaining basic data for the meta-analysis, several methodological difficulties, all of which could have been corrected through better report-writing and publishing procedures, emerged.

The first major problem encountered involved the unavailability of statistical data necessary for effect size calculations. Many studies had to be excluded from the meta-analysis because data was not reported

or reported incompletely. Perhaps the most glaring examples of failures to report useful data were the many studies which presented only significant statistical results. Frequently non-significant results were mentioned only casually or dismissed briefly on statistical tables with a dash or a n.s. with no indication of the magnitude of the statistical test or probability level. On several occasions it was difficult to determine whether or not a test for sex differences in activity level promised in the abstract had even been conducted. Consequently, for many otherwise relevant studies effect sizes could not be calculated at all. In the interests of more complete future meta-analyses, Glass (1977) suggests that journal contributors and editors ensure that all meaningful original data be accessible, whether in published or unpublished form, for research integration.

An additional methodological problem (reflected in the large Unknown categories in Table II) concerns many researchers' incomplete reporting of study characteristics which could have contributed raw data for a correlational analysis to assess their influence on sex differences in activity level. For example, in a number of studies, children's activity level was measured in peer groups, but the authors neglected to mention the number of children in the group and whether they were the same sex or a different sex from the target child. Other studies were unclear as to whether or not a parent or teacher was present or absent during the measurement process. In still other studies, it was impossible to determine how many raters rated the subjects, whether the rater was a teacher, parent or researcher and even how many data points were involved in the rating.

A third and related problem involved the frequent failure on the part of researchers to report complete reliability data for their activity level measure. Many researchers seemed to have neglected the reliability issue altogether. Many of those who were concerned with it reported reliability in a confused, imprecise fashion, making it difficult to determine which reliability coefficient applied to which dependent measure, which kind of reliability was being assessed, etc. Consequently valuable information on the relationship between sex differences in activity level and measurement reliability remained inaccessible.

In summary, a sex difference of .52 of a standard deviation is evaluated as moderate in size. Results of the meta-analysis indicate that sex accounts for only 5.9% of the total variance in activity level; clearly, the distributions of the two sexes show considerable overlap. The finding that sex differences exist has no implications for etiology and it is impossible to say whether sex differences are innate or produced through socialization. The social implications of a moderately sized sex difference in gross motor activity level may, however, be greater than size alone suggests. For example, a gender difference in activity level has clear implications for the development of exaggerated sex-role stereotypes. As Hyde (1981) points out, although overall mean differences in the population are small, large differences may occur at the tails of distributions. Hyde observes that assuming a gender difference of .40 standard deviation and a cut-off point for a given behaviour or ability at the 95th percentile, 7.35% of males will be above the cutoff as opposed to 3.22% of females. This amounts to a 2 to

l ratio of males to females. The higher the cutoff point, the larger the ratio becomes. When applied to sex differences in activity level, this observation may help to explain sex-role stereotypes which automatically characterize males as the more active sex. Nevertheless, although a rationale for the social stereotype exists, the stereotype exaggerates the actual magnitude of the difference in the total population out of all proportion. Hyperactivity research (Tieger, 1980; Ross & Pelham, 1981) suggests that high activity level occurs in conjunction with other so-called problem behaviours such as low attention span and high aggressiveness. In a setting where good attention span and behavioural control are important prerequisites (e.g. school), then the more motorically active males are far more likely to be noticed and stereotyped as overactive than the less active females.

In conclusion, results of the meta-analysis suggest that Maccoby and Jacklin's dismissal of gross motor activity level as a possible basis for sex-differential responding calls for re-evaluation. Data from the meta-analysis, which unlike the box-score approach used by Maccoby and Jacklin, quantitatively summarizes comparisons from many different studies, points to the existence of a sex difference of moderate size in gross motor activity level. These findings are robust in that they do not seem to be related to such measurement issues as reliability, type of measure, obtrusiveness, inclusiveness, number of raters/observers, number of data points or length of behavioural sample. Neither do the results seem to be affected by researcher bias as measured by sex of researcher or by the bias to confirm major hypotheses. Data from the meta-analysis are inconclusive on several of

the points raised by Maccoby and Jacklin (age patterns, peer presence and setting characteristics such as novelty, stressfulness, and restrictiveness) and these hypotheses clearly require further investigation. The number of comparisons summarized, the magnitude of the average effect size and the robustness of the findings, however, all argue strongly for the existence of a stable sex difference in gross motor activity level. If, as the results suggest, such a sex difference exists, it could constitute what Bell and Harper (1977) call a "child effect," a behavioural cue emanating from the child which elicits a parental response that contributes to the differential socialization of males and females.

Reference Notes

1. Eaton, W., Nottelmann, E., Williams, J., and Williams, K. Behavior in structured and informal classrooms. Unpublished manuscript, University of Manitoba, 1978.

References

Bell, R., & Harper, L. Child effects on adults. Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers, 1977.

Block, J. Issues, problems, and pitfalls in assessing sex differences: A critical review of The Psychology of Sex Differences. Merrill-Palmer Quarterly, 1976, 22, 283-308.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Statistical power analysis for the behavioral sciences. (2nd. ed.). New York: Academic Press, 1969.

Cooper, H. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 1979, 37, 131-146.

Cooper, H., Burger, J., & Good, T. Gender differences in the academic locus of control beliefs of young children. Journal of Personality and Social Psychology, 1981, 40, 562-572.

Eagly, A., & Carli, L. Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. Psychological Bulletin. 1981, 90, 1-20.

Eaton, W.O. Measuring activity level with actometers: Reliability, validity, and arm length. Child Development, 1983, in press.

Eaton, W.O., & Keats, J.G. Peer presence, stress, and sex differences in the motor activity levels of preschoolers. Developmental Psychology, 1982, 18, in press.

Epstein, S. The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and Social Psychology, 1979, 37, 1097-1126.

Eysenck, H. An exercise in mega-silliness. American Psychologist, 1978,33, 517. (Comment)

Gallo, P. Meta-analysis--a mixed meta-phor? American Psychologist, 1978, 33, 515-517. (Comment)

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

Glass, G. Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.), Review of Research in Education, 1977, 5, 351-379.

Glass, G. Summarizing effect sizes. New Directions for Methodology of Social and Behavioral Science, 1980, 5, 13-31.

Glass, G., & Smith, M. Reply to Eysenck. American Psychologist, 1978, 33, 515-519. (Comment)

Hall, J. Gender effects in decoding nonverbal cues. Psychological Bulletin, 1978, 85, 845-857.

Hyde, J.S. How large are cognitive gender differences? American Psychologist, 1981, 36, 892-901.

Light, R., & Smith, P. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Maccoby, E., & Feldman, S. Mother-attachment and stranger-reactions in the third year of life. Monographs of the Society for Research in Child Development, 1972, 37, 1-63.

Maccoby, E., & Jacklin, C. The psychology of sex differences. Stanford, Ca.: Stanford University Press, 1974.

Maccoby, E.E., & Jacklin, C. Sex differences in aggression: A rejoinder and reprise. Child Development, 1980, 51, 964-980.

Pedersen, F., & Bell, R. Sex differences in preschool children without histories of complications of pregnancy and delivery. Developmental Psychology, 1970, 3, 10-15.

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

Rosenthal, R. The 'file drawer problem' and tolerance for null results. Psychological Bulletin, 1979, 86, 638-641.

Rosenthal, R. Summarizing significance levels. New Directions for Methodology of Social and Behavioral Science, 1980, 5, 33-46.

Ross, A., & Pelham, W. Child psychopathology. Annual Review of Psychology, 1981 32, 243-278.

Serbin, L., O'Leary, K., Kent, R., & Tonick, I. A comparison of teacher response to the pre-academic and problem behavior of boys and girls. Child Development, 1973, 44, 796-804.

Smith, M., & Glass, G. Meta-analysis of psychotherapy outcomes. American Psychologist, 1977, 32, 752-760.

Smith, M., Glass, G., & Miller, T. The benefits of psychotherapy. Baltimore: The John Hopkins University Press, 1980.

Smith, P.K., & Daglish, L. Sex differences in parent and infant behavior in the home. Child Development, 1977, 48, 1250-1254.

Tieger, T. On the biological basis of sex differences in aggression. Child Development, 1980, 51, 943-963.

Table I

Citation, ES and Estimated Z of Studies

Included in Activity Level Meta-Analysis

Citation	Measure Type	ES	ES N	Mean ES	Mean ES N	Est. Z
Achenbach, 1969	R	1.26	32			3.02
Ault et al., 1972	R	2.27	25			3.75
Battle & Lacey, 1972	R	.04	64	.17	60	.66
	R	.15	61			
	R	.32	54			
Bell et al., 1971	M	-.10	74	.45	74	1.89
	M	.58	74			
	O	1.86	74			
	O	.49	74			
	R	.04	74			
	O	.60	74			
	O	.07	74			
	O	.07	74			
Bjorklund & Butter, 1973	R	.37	132			2.09
Bronson, 1966	R	-.28	85	.09	85	.41
	R	.20	85			
	R	.22	85			
	R	.22	85			
Buss et al., 1980	M	.24	129	.29	129	1.63
	M	.41	129			
	R	.29	129			
	R	.44	129			
	R	.05	129			
Crowther et al., 1981	R	-.20	25			-.50
	R	.32	106			1.63
	R	.59	212			4.12
	R	.47	245			3.58
DiPietro, 1981	R	.56	52	.39	49	1.34
	R	.17	48			
	R	.45	47			

Table I (continued)

Citation	Measure Type	ES	ES N	Mean ES	Mean ES N	Est. Z
Elder, 1970	R	.44	27			1.12
Fales, 1937	R	.14	32			.40
Feiring & Lewis, 1980	R	.49	60	.12	59	.46
	R	-.20	62			
	R	.07	62			
	R	.12	53			
	O	-.01	60			
	O	.32	62			
	M	.08	54			
Garside et al., 1975	R	.07	209			.51
Goggin, 1975	O	.47	73			2.00
Goodenough, 1930	R	.40	16			.78
	R	1.89	17			2.83
Halverson & Waldrop, 1973	M	1.33	58			4.22
	M	.02	59	.48	59	1.79
	M	.94	59			
Harrison, 1941	O	.92	40			2.64
Hattwick, 1937	R	.12	579	.20	579	2.39
	R	.27	579			
Kaspar et al., 1971	M	.86	36	.56	36	1.62
	M	.98	36			
	M	-.02	36			
	M	.40	36			
Kurtz, 1969	R	.34	169			2.18
Kurtz, 1971	R	1.07	40			2.98
Lahey et al., 1980	R	1.12	42			3.17
	R	.06	41			.19
Loo & Wenar, 1971	M	.14	40			.44

Table I (continued)

Citation	Measure Type	ES	ES N	Mean ES	Mean ES N	Est. Z
MacFarlane et al., 1962	R	.40	116	.39	116	2.06
	R	.11	98			
	R	.08	88			
	R	.24	94			
	R	.29	91			
	R	.09	83			
	R	.49	83			
	R	.69	75			
	R	.47	77			
	R	.39	61			
	R	.47	65			
	R	1.40	65			
	R	1.37	58			
	R	1.10	41			
	R	.39	116			
Melson, 1977	O	3.09	34	1.23	34	3.06
	O	2.40	34			
	O	.41	34			
	O	.08	34			
	O	.18	34			
Moss, 1967	O	.62	29	.43	27	1.09
	O	.23	25			
Nelson, 1931	O	.14	91			.67
Paulsen & Johnson, 1980	R	.91	55	.85	55	2.90
	R	.78	55			
Richman et al., 1975	R	.30	657			3.80
Rose & Mayer, 1968	M	.10	29			.27
Rowe & Plomin, 1977	R	.59	182			3.82
Seifer et al., 1981	R	.43	309			3.70
Smith & Daglish, 1977	O	.91	16			1.66
	O	.65	16			1.24

Table I (continued)

Citation	Measure Type	ES	ES N	Mean ES	Mean ES N	Est. Z
Spring et al., 1977	R	.46	191			3.10
	R	.55	276			4.40
	R	.36	193			2.46
	R	.79	259			5.91
	R	.52	221			3.74
Stein & Lenrow, 1970	R	.49	249			3.76
Stone, 1981	R	.48	25000	.38	25000	29.51
	R	.27	25000			
Tauber, 1979	O	.39	146			2.31
Walker, 1967	R	.74	406			6.99
Willerman, 1973	R	.58	54			2.05
	R	-.16	39			-.50
Willerman & Plomin, 1973	R	.06	86			.28
	R	.05	43			.16
Wolfensberger et al., 1962	M	.71	100			3.35

Note: Mean ES and Mean ES sample size were calculated for each study yielding non-independent effect sizes. For measure type: R=Rating, M=Mechanical, O=Observational.

Table II

Reliability Coefficients for All Categorical Variables

Variable	Kappa	Spearman-Brown Corrected
Source	.73	.84
Test	.44	.61
Basis for selection	.59	.74
Type of setting	.77	.87
Peer presence	.79	.88
Peer sex (same or mixed)	.80	.89
Adult presence	.46	.63
Novelty of setting	.71	.83
Stressfulness of setting	.47	.64
Restrictiveness of setting	.58	.73
Obtrusivenss of measure	.90	.95
Inclusiveness of measure	.52	.69
Type of measure	.88	.94
Description of rater	.90	.95
Type of reliability coefficient	.77	.87
Sufficient information to calculate ES	1.00	1.00
Direction of effect	.87	.93
Significance of results	.89	.94
Exactness of probability estimate	.65	.79
Statistic used to calculate ES	.98	.99
Type of error term used for calculations	.81	.90

Table III

Reliability Coefficients for All Continuous Variables

Variable	Inter-Rater	Spearman-Brown Corrected
Total N	1.00	1.00
No. of males	.99	1.00
No. of females	1.00	1.00
Mean/median age of subjects	.99	1.00
Age range of subjects:		
Minimum	1.00	1.00
Maximum	1.00	1.00
Peer no.	1.00	1.00
No. of scale points per item (for rating scales)	.47	.64
No. of items on scale (for rating scales)	.13	.23
No. of raters	.76	.86
No. of observers	.47	.64
Length of behavioural sample in minutes	.98	.99
Reliability coefficient	.76	.86
Probability	.91	.95

Table IV

Mean ES, Median ES and Correlations with Log ES for all Categorical Variables

Variable	N	Mean ES	Median ES	r with log ES	p
Test central to hypothesis	104			-.03	.76
Yes	25	.52	.32		
No	79	.48	.40		
Source	104			.07	.49
Journal	82	.48	.40		
Book	22	.52	.40		
Other					
Basis for selection	99			-.20	.05
Public School	14	.59	.54		
Nursery School/ Daycare	23	.62	.40		
Summer School/Camp	1	2.27	2.27		
Community Organization	2	.27	.27		
Other	59	.38	.32		
Unknown	5				
Type of setting	102			-.12	.21
Preschool	40	.56	.40		
School Classroom	11	.70	.52		
Lab	9	.35	.39		
Home	29	.43	.39		
Other	13	.38	.32		
Unknown	2				
Peer presence	69			.13	.28
Absent	16	.40	.40		
Present	53	.58	.41		
Unknown	35				
If present,	53			.01	.96
Same-sex	11	.59	.56		
Mixed-sex	42	.58	.41		
Unknown	51				
Adult presence	96			.02	.81
Absent	2	.29	.29		
Present	94	.49	.39		
Unknown	8				

Table IV (continued)

Variable	N	Mean ES	Median ES	r with log ES	p
Novelty of setting	94			-.10	.36
Low	82	.51	.40		
Medium & High	12	.33	.36		
Unknown	10				
Stressfulness of setting	94			-.06	.58
Low	72	.52	.41		
Medium & High	22	.47	.38		
Unknown	10				
Restrictiveness of setting	95			-.17	.11
Low	55	.57	.41		
Medium	33	.44	.37		
High	7	.30	.14		
Unknown	9				
Obtrusiveness of measure	103			.02	.80
Unobtrusive	79	.50	.39		
Mildly obtrusive	17	.51	.44		
Very obtrusive	7	.40	.40		
Unknown	1				
Inclusiveness of measure	103			-.09	.36
Low	24	.62	.40		
Moderate	29	.50	.41		
High	50	.44	.39		
Unknown	1				
Type of measure	104			.16	.12
Rating scale	66	.43	.38		
Mechanical	15	.44	.40		
Observational	23	.71	.41		
Other	0				
Description of rater	69			-.21	.09
Self	5	.54	.49		
Parent	21	.47	.40		
Teacher	21	.52	.44		
Investigator	19	.35	.22		
Other	3	.17	.15		
Unknown	35				

Table IV (continued)

Variable	N	Mean ES	Median ES	r with log ES	p
Type of reliability coefficient	58			-.14	.31
Test-retest	10	.52	.54		
Split half (alpha)	12	.50	.38		
Observer agreement	36	.48	.32		
Unknown	46				
Sufficient information to calculate ES	104			1.00	.00
Yes	104	.49	.40		
No	0				
Direction of effect	103			-.64	.0001
M>F	96	.54	.41		
F>M	7	-.14	-.16		
Unknown	1				
Significance of results by author's standard	91			-.55	.0001
Significant	43	.82	.59		
Non-significant	48	.28	.18		
Unknown	13				
Exactness of probability estimate	36			.37	.03
Exact	3	.24	.27		
Inexact	33	.84	.59		
Unknown	68				
Statistic used to calculate ES	104			-.29	.003
F	0				
t	26	.75	.45		
r	3	.61	.78		
proportions	17	.53	.40		
means	58	.36	.32		
Type of error term used to calculate ES	89			-.15	.17
Pooled estimate from study	50	.55	.42		
Own pooled estimate	39	.39	.32		
Within group male	0				
Within group female	0				
Unknown	15				

Table IV (continued)

Variable	N	Mean ES	Median ES	r with log ES	p
%age male authorship	104			.03	.74
0%	41	.54	.39		
33%	4	.30	.40		
50%	13	.37	.14		
67%	14	.52	.35		
75%	2	.59	.59		
100%	28	.50	.48		
Unknown	2	.19	.19		

Note: N does not total 104 for studies in which the predictor variable could not be coded.

Table V

Correlations of Continuous Variables with Log ES

Variable	N	r with log ES	p
Year of publication	104	-.06	.57
Total N	104	.01	.94
Number of males	104	.01	.94
Number of females	104	.01	.94
Mean/median age of subjects	104	.18	.06
Age range of subjects:			
Minimum	96	.29	.003
Maximum	96	.23	.03
Number of peers present	12	.42	.17
Number of scale points per item (for rating scales)	69	-.01	.91
Number of items on rating scale	68	-.01	.96
Number of raters	55	-.16	.25
Number of observers	20	-.04	.86
Length of behavioural sample	94	-.03	.84
Reliability coefficient	57	-.03	.84
Probability	37	-.47	.003

Table VI

Comparison with Other Sex Difference Meta-Analyses

Variable	Mean	(N)	Source
Gross Motor Activity Level	.52	(54)	
Preschooler Aggressiveness	.63	(25)	Maccoby & Jacklin, 1980
Persuasion Studies	-.16	(33)	Eagly & Carli, 1981
Group Pressure Conformity	-.32	(46)	Eagly & Carli, 1981
Other conformity	-.28	(11)	Eagly & Carli, 1981
Decoding Nonverbal Cues	-.25	(25)	Hall, 1978
Verbal Ability	-.35	(12)	Hyde, 1981
Visual-Spatial Ability	.47	(7)	Hyde, 1981
Field Articulation	.55	(14)	Hyde, 1981
Academic Locus of Control	-.10	(10)	Cooper et al., 1981

Note: Means based only on studies where effect size estimates were calculated.

Appendix A

Citations of Studies Used for Meta-Analysis of Sex Differences in Activity Level

Achenbach, T.M. Cue learning, associative responding, and school performance in children. Developmental Psychology, 1969, 1, 717-725.

Ault, R.L., Crawford, D.E., & Jeffrey, W.E. Visual scanning strategies of reflective, impulsive, fast-accurate, and slow-inaccurate children on the Matching Familiar Figures Test. Child Development, 1972, 43, 1412-1417.

Battle, E.S., & Lacey, B. A context for hyperactivity in children over time. Child Development, 1972, 43, 757-753.

Bell, R.Q., Weller, G.M., & Waldrop, M.F. Newborn and preschooler organization of behavior and relations between periods. Monographs of the Society for Research in Child Development, 1971, 36, Series No. 142.

Bjorklund, D.F., & Butter, E.J. Can cognitive impulsivity be predicted from classroom behavior? Journal of Genetic Psychology, 1973, 123, 185-194.

Bronson, W. Central orientations: A study of behavior organization from childhood to adolescence. Child Development, 1966, 37, 125-155.

Buss, D.M., Block, J.H., & Block, J. Preschool activity level: Personality correlates and developmental implications. Child Development, 1980, 51, 401-408.

Crowther, J.H., Bond, L.A., & Rolf, J.E. The incidence, prevalence and severity of behavior disorders among preschool-aged children in day care. Journal of Abnormal Child Psychology, 1981, 9, 23-42.

DiPietro, J. Rough and tumble play: A function of gender. Developmental Psychology, 1981, 17, 50-58.

Elder, M. S. The effects of temperature and position on the sucking pressure of newborn infants. Child Development, 1970, 41, 95-102.

Fales, E. A comparison of the vigorousness of play activities of preschool boys and girls. Child Development, 1937, 8, 144-158.

Feiring, C. & Lewis, M. Temperament: Sex differences and stability in vigor, activity, and persistence in the first three years of life. Journal of Genetic Psychology, 1980, 136, 65-75.

Garside, R.F., Birch, H., Scott, D.M., Chambers, S., Kolvin, I., Tweddle, E.G., & Barber, L.M. Dimensions of temperament in infant school children. Journal of Child Psychology and Psychiatry, 1975, 16, 219-231.

Goggin, J. Sex differences in the activity level of preschool children as a possible precursor of hyperactivity. Journal of Genetic Psychology, 1975, 127, 75-81.

Goodenough, F.L. Inter-relationships in the behaviour of young children. Child Development, 1930, 1, 29-47.

Halverson, C., & Waldrop, M. The relations of mechanically recorded activity level to varieties of preschool behavior. Child Development, 1973, 44, 678-681.

Harrison, R. Personal tempo and the interrelationship of voluntary and maximal rates of movement. Journal of General Psychology, 1941, 24, 343-379.

Hattwick, L.A. Sex differences in behavior of nursery school children. Child Development, 1937, 8, 343-355.

Kaspar, J.C., Millichamp, J.G., Backus, R., Child, D., & Schulman, J.L. A study of the relationship between neurological evidence of brain damage in children and activity and distractibility. Journal of Consulting and Clinical Psychology, 1971, 36, 329-337.

Kurtz, R. Sex differences and variations in body attitudes. Journal of Consulting and Clinical Psychology, 1969, 33, 625-629.

Kurtz, R.M. Body attitude and self-esteem. Proceedings of the 79th Annual Convention of the APA, 1971, 8, 467-468.

Lahey, B., Hammer, D., Crumrine, P., & Forehand, R. Birth order X sex interactions in child behavior problems. Developmental Psychology, 1980, 16, 608-615.

Loo, C., & Wenar, C. Activity level and motor inhibition: Their relationship to intelligence-test performance in normal children. Child Development, 1971, 42, 967-971.

MacFarlane, J.W., Allen,L., & Honzik, M.P. A developmental study of the behavior problems of normal children between 21 months and 14 years. Berkeley: University of California Press, 1962.

Melson, G. Sex differences in use of indoor space by preschool children. Perceptual and Motor Skills, 1977, 44, 207-213.

Moss, H.A. Early sex differences and mother-infant interaction. In R.C. Friedman, R.N. Richart, & R.L. Vande Weile (Eds.), Sex differences in behavior. New York: John Wiley & Sons, 1974.

Nelson, J. Personality and intelligence. New York: Columbia Teachers College, 1931.

Paulsen, K., & Johnson, M. Impulsivity: A multidimensional concept with developmental aspects. Journal of Abnormal Child Psychology, 1980, 8, 269-277.

Richman, N., Stevenson, J.F., & Graham, P.J. Prevalence of behavior problems in 3-year-old children: An epidemiological study in a London borough. Journal of Child Psychology and Psychiatry, 1975, 16, 277-287.

Rose, H.E., & Mayer, J. Activity, calorie intake, fat storage and the energy expenditure balance of infants. Pediatrics, 1968, 41, 18-29.

Rowe, D.C., & Plomin, R. Temperament in early childhood. Journal of Personality Assessment, 1977, 41, 150-156.

Seifer, R., Sameroff, A., and Jones, F. Adaptive behavior in young children of emotionally disturbed children. Journal of Applied Developmental Psychology, 1981, 1, 251-276.

Smith, P.K., & Daglish, L. Sex differences in parent and infant behavior in the home. Child Development, 1977, 48, 1250-1254.

Spring, C., Blunden, D., Greenberg, L.M. & Yellin, A.M. Validity and norms of a hyperactivity rating scale. Journal of Special Education, 1977, 11, 313-321.

Stein, K.B., & Lenrow, P. Expressive styles and their measurement. Journal of Personality and Social Psychology, 1970, 16, 656-664.

Stone, F.B. Behavior problems of elementary-school children. Journal of Abnormal Child Psychology, 1981, 9, 407-418.

Tauber, M. Parental socialization techniques and sex differences in children's play. Child Development, 1979, 50, 225-234.

Walker, R.N. Some temperament traits in children as viewed by their peers, their teachers, and themselves. Monographs of the Society for Research in Child Development, 1967, 32, Serial No. 114, 1-36.

Willerman, L. Activity level and hyperactivity in twins. Child Development, 1973, 44, 188-293.

Willerman, L., & Plomin, R. Activity level in children and their parents. Child Development, 1973, 44, 854-858.

Wolfensberger, W., Miller, M., Foshee, J., & Cromwell, R. Rorschach correlates of activity level in high school children. Journal of Consulting Psychology, 1962, 26, 269-272.

Appendix B

Meta-Analysis of Sex Differences in Activity Level

Coding Categories and Definitions

I. General Information

Source: Journal, Book or Other.

Test of sex differences in activity level is central to hypothesis: Yes--the study focuses on activity level and sex differences in activity level constitutes one of the main hypotheses. No--sex differences in activity level were not hypothesed and were analyzed only incidentally. e.g. study has no directional hypothesis regarding sex differences.

Total number of comparisons this study: Number of tests for sex differences in activity level conducted in this study.

Number of this comparison: If there are multiple tests for sex differences in activity level, this is the number of the particular comparison being coded, e.g., first of three comparisons.

Measure number: If more than one dependent measure of activity level is employed, state the number and type of the measure used.

II. Subject Characteristics

Total N: For longitudinal studies, this is the maximum sample size tested at any one time.

No. of males: For longitudinal studies, this is the maximum no. of males tested at any one time.

No. of females: For longitudinal studies, this is the maximum no. of females tested at any one time.

Mean/median age: If age range only is given, use mean of the age range. e.g. age range of 3 years (0 months - 3 months): Mean/median will be 18 months (1-6). If subjects are "8-year olds," mean age is 102 months (8-6); minimum age is 96 months (8-0); and maximum age is 107 months (8-11). If subjects are in e.g. Grade 3, then mean, minimum and maximum ages are the same as for 8-year olds; Grade 1 children are treated as 6-year olds, etc. "College age subjects" are assumed to be in the 216 month (18-0) to 252 month (21-11) age range.

Age range: If subjects are all the same age, minimum, maximum and mean/median will all be the same.

Basis of selection: Public school--subjects are children selected from grades one to twelve or thirteen at school. Nursery school/day care--subjects are preschool children selected from daycare, nursery school or kindergarten. Summer school/camp--subjects are children selected from summer school or day camps operated during the summer. Community organization--subjects have been selected from some institutions other than the ones listed above e.g., hospital wards, medical clinics, community associations, clubs, etc. Other--subjects selected in some fashion not included in alternatives described above.

III. Characteristics of Setting

Type of Setting: Preschool--Subjects are children whose activity level was measured in kindergarten, nursery school or daycare setting. School classroom--Subjects are children whose activity level was measured in a school classroom. Lab--Subjects' activity level was measured in a contrived setting such as a laboratory or room purposely set up for the study. Home--Subjects' activity level was measured in their homes under naturalistic conditions. Other--Subjects' activity level measured in a setting other than the ones described above. All categories--If activity level is measured across more than one setting, code the modal setting for the activity level measure when it is possible to determine the main or modal setting in which activity level is measured e.g. if a mother rates her child for overall activity level and she most frequently observes him at home, code "home." If activity level is measured in more than one setting, and it is impossible to determine the modal setting, code "other."

Peer presence: Absent--Activity level measured while peers are not present. Present--Peers are present. If activity level is measured across more than one setting, code whether peers are present or absent and whether they are same or mixed-sex in the modal setting. e.g. If school aged children are being observed during a routine day, presence of mixed-sex peers would be coded.

If subjects are children (12 years or younger), adult presence: Absent--Adults are not present. Present--Adults are present.

Novelty of setting: Setting encountered frequently--Setting is one which subject encounters frequently e.g., home, school classroom for a child. S has been introduced to setting before--Setting is one which subject has been in before e.g., lab setting with which S has been familiarized, etc. S has never been in setting before--S has never seen setting in which activity level is measured. All categories: If activity level is measured across more than one setting, code novelty of the modal setting.

Stressfulness of setting: Low--Setting is one which should not cause subject to be under stress e.g., it is familiar to him, involves no complicated apparatus, etc. such as free play in a home setting Medium--Setting is one which should cause subject only an average amount of stress e.g., apparatus is present but not obtrusive, etc., such as in a school classroom. High--Setting is one which causes S considerable stress, e.g. he is being tested and timed, apparatus is very obtrusive, etc. All categories: If activity level is measured in more than one setting, code stressfulness of the modal setting.

Restrictiveness of setting: Low--setting imposes few motor restrictions on subjects e.g. usually true for a child when activity level is measured during free play or in an unspecified home setting. Medium--setting imposes a moderate number of motor restrictions e.g. usually true for a child when activity level is measured during school classtime or during regular preschool routine. High--setting is very restrictive motorically e.g., activity level is measured in a small room with limited number of things to do. All categories: If activity level is measured in more than one setting, code restrictiveness of the modal setting. If it is impossible to determine the modal setting, code "Unknown."

Obtrusiveness of measure: Unobtrusive--e.g., activity level measured by hidden rater, by teacher in a classroom setting, through self-ratings, or by some measure which does not affect the subject's activity level behaviour. Ordinary observational methods without elaborate apparatus fall into this category. Mildly obtrusive--e.g., activity level is measured by a strange rater who is not hidden and has eye-catching apparatus. which may mildly affect the subject's behaviour. Actometers usually fall under this category. Very obtrusive--e.g. complicated apparatus is used which the subject is forced to notice (crank-turning, ballistograph, etc.) and which probably affects his behaviour.

IV. Type of Measure

Inclusiveness of activity level measure: Low--Measure is specific in that it involves one highly specific measure hypothesized to be indicative of general activity level e.g. arm movement. For a rating scale, one very specific item or several highly related and specific items are used. Moderate--measure includes several specific measures of activity level (e.g. movement of arms and legs. High--measure of activity level is based on a broad and inclusive sample of subject's behaviour e.g., teacher ratings of children's activity level in the classroom. For a rating scale, general and multiple items are used.

Type of measure: Rating scale--measure is rating scale based on subjective impressions of subjects' activity level. -- includes self rating scales, teacher rating scales, parent rating scales, and many observational rating scales. Mechanical--measure involves mechanical apparatus e.g., actometers, stabilometers, ballistographs. Observational--measure is estimate of activity level based on actual observations of behaviour. May include use of observational rating scales. Other--any measure not included above.

A. Rating Scales

Number of scale points per item: For a rating scale, refers to the highest number a subject may be assigned on one particular item of the scale e.g., If the item is "how active is your child?" and the child is assigned a number from 1 to 7, number of scale points per item would be 7. The no. of scale points used for calculating the statistic on which the effect size is based is used for this coding category. Items that are dropped later during data analysis are still included in this estimate.

Number of items on scale: Refers to total number of questions or items assessing activity level.

C. All Measures

Length of behavioural sample in minutes: For all measures, refers to total number of minutes during which activity level is measured. If more than one rater or observer is used, sum across raters or observers e.g. if 2 raters observed for 15 minutes, this item is coded as 30 minutes. Where a sample length range is given, code the mean of the range.

Type of reliability coefficient: Test-retest reliability estimated by correlation between two administrations of some measure or between parallel forms of measure. Split half or alpha--measure's reliability is estimated by intercorrelation of items or components. Observer agreement--measure's reliability estimated by correlation of independent observations of two observers. Other--measure's reliability is estimated by a combination of the types of reliability coefficients described above.

Summary of author's implicit/explicit definition of activity level: Brief summary of original author's operational definition of activity level as used in study e.g., activity level is defined as amount of gross motor activity in free play situation.

V. Calculations of Effect Size

Sufficient information to calculate effect size (ES): Yes--study in published form contains statistics necessary to calculate or estimate an effect size for activity level. No--Inadequate information.

Direction of effect: M>F--male activity level is greater than female. F>M--female activity level is greater than male. Unknown--study does not indicate whether male or female activity level is greater.

Significance of results by author's standards: Significant--study reports that males and females have significantly different ($p<.05$) activity levels. Non-significant--study reports that males and females do not have significantly different ($p>.05$) activity levels.

Alpha probability: Alpha probability level of results obtained for a particular measure as reported by authors of original study.

Exactness of probability estimate: Exact--alpha probability level of results obtained for a particular measure are reported exactly by authors of original study e.g., $p=.03$ not $p<.05$. Inexact--alpha probability level of results obtained for a particular measure are not reported exactly by authors of original study e.g., $p<.05$ not $p=.041$.

Statistic used to calculate ES: Refers to statistic from study that is used to calculate or estimate effect size for meta-analysis, e.g., t-statistic, ANOVA table, etc.

Type of error term used for calculations: Pooled estimate from study--male and female pooled error term from study was used to calculate effect size for the meta-analysis. Own pooled-estimate--a weighted pooled error term devised for this meta-analysis was used to calculate the effect size. Within group male--male error term reported in study used to calculate effect size. Within group female--female error term reported in study used to calculate effect size.

Effect size: Effect size calculated for this particular measure, e.g., ES=.37.

Appendix C

Meta-Analysis of Sex Differences in Activity Level

Coding Form

I. General Information

Date coded _____

Study No....._____

Citation _____

Coder....._____

Year of Publication....._____

Source: Journal, Book or Other....._____

Test of sex differences in activity level is central to hypothesis: Yes or No....._____

Total no. of comparisons this study....._____

No. of this comparison....._____

Measure No....._____

Describe _____

II. Subject Characteristics

Total N....._____

No. of males (.) unknown....._____

No. of females (.) unknown....._____

(If n's are missing but total N is reported, assume equal n for both sexes)

Mean/median age of subjects....._____

Age range of subjects: minimum....._____

maximum....._____

Basis for selection: (1)public school (2)nursery school/ day care (3)summer school/camp (4)community organization e.g. hospital (5)other: describe_____ ..____

III. Characteristics of Setting

Type of setting: (1)preschool (2)school classroom (3)lab (4)home (5)other: describe_____

_____ .....____

Peer presence: (1)absent (2)present.....____
If present, approximate number.....____
If present, (1)same-sex (2)mixed sex.....____

If subjects are children, adult presence (1)absent (2)present.....____

Novelty of setting: (1)setting encountered frequently (2)S has been introduced to setting before (3)S has never been in setting before (.)unknown.....____

Stressfulness of setting: (1)low (2)medium (3)high (.)unknown.....____

Restrictiveness of setting: (1)low (2)medium (3)high (.)unknown.....____

Obtrusiveness of measure: (1)unobtrusive (2)mildly obtrusive (3)very obtrusive (.)unknown.....____

IV. Type of Measure

Inclusiveness of activity level measure (1)low (measure is specific) (2)moderate (3)high (measure is broad and inclusive) (.)unknown.....____

Type of measure (1)rating scale (2)mechanical: describe: _____ (3)observational:_____ describe:_____ (4)other: describe:_____ _____.....

A. Rating Scales

No. of scale points per item in rating scale (.)unknown/ not applicable.....

No. of items on scale (.)unknown/not applicable.....

No. of raters (.)unknown/not applicable.....

Description of rater: (1)self (2)parent (3)teacher (4)investigator (5)other: describe_____ _____.....

B. Observational Measures

No. of observers.....

C. All Measures

Length of behavioral sample in minutes (.)unknown/not applicable (999)very large sample of unknown length e.g. self rating.....

Reliability coefficient (.)unknown.....

Type of reliability coefficient (1)test-retest (2)split half or alpha (3)observer agreement (.)unknown.....

Summary of author's implicit/explicit definition of activity level_____

_____

V. Calculations of Effect Size

Sufficient information to calculate ES: Yes or No.....

Direction of effect: (1)M F (2)F M (.)unknown.....

Significance of results by author's standard (1)significant (2)non-significant Note if different from .05 two-tailed

_____.....____

Probability (.)unknown.....____

Exactness of probability estimate (2)exact (2)inexact....____

Mean of males (.)unknown.....____

Standard deviation of males (.)unknown.....____

Mean of females (.)unknown.....____

Standard deviation of females (.)unknown.....____

Statistic used to calculate ES: (1)F (2)t (3)r (4)proportions (5)means and standard deviations (6)other: describe:_____.....____

Type of error term used for calculations: (1)pooled estimate from study (2)own pooled estimate (3)within group male (4)within group female.....____

Effect size.....____