Journal of Gerontology
1981, Vol. 36, No. 5, 634-638

Letters to the Editor

Editorial Comment:

The two letters that follow were invited for publication in the *Journal of Gerontology* by a former associate editor (Larry Thompson) during his tenure in office. This action was precipitated by the controversy that arose during the review process of the manuscript entitled "Modifiability of Fluid Intelligence in Aging: A Training Approach" by J. K. Plemons, S. L. Willis, and P. B. Baltes and published in the *Journal of Gerontology* in 1978. A number of the reviewers felt that the research described in the paper was only mildly supportive of the hypotheses, that the effects were overgeneralized and not supportive of several conclusions reached by the authors, and that important methodological flaws were overlooked. Still others felt that the import of the paper overshadowed the problems and that it should be published.

Because at that time the issues touched upon in the study were informative and controversial, it was decided that the paper would be of interest to the readers and that they should be given the opportunity to evaluate the evidence for themselves. However, it was also agreed that the paper should be followed by a critique from one of the opposing reviewers. Dr. Gary Donaldson's critique was the most comprehensive of those submitted. Therefore, he was invited to submit a brief statement incorporating the basic ideas of the critique. To insure fairness Plemons et al. were given the opportunity to reply to Donaldson's criticisms.

Unfortunately a number of unforeseen complications has delayed the timely completion of this proposed plan. The readers may feel that it is inappropriate to revive this controversy at such a late date. However, the editorial staff made a commitment to the authors and reviewers that their positions would be aired. Now, all the pieces are in place and the commitment can be fulfilled.

It is hoped that the readers will profit from the arguments posed in the critique and rebuttal. Perhaps the perspective of time itself will lend clarity and objectivity to the arguments, and the controversy can be laid to rest for some time to come. At the very least these discussions should remind us of the importance of methodological precision in our branch of the sciences.

Larry W. Thompson

Dear Editor:

The research described by Plemons, Willis, and Baltes (1978) concerns the modification of an ability, fluid intelligence, by direct experimental intervention. This is an intriguing topic, worthy of well-designed research; unfortunately, problems of design impede acceptance of the authors' contention that genuine change in intellectual ability has been observed in this study. It is my intent neither to denigrate intervention research in general, nor to provide a defense of the theory of fluid and crystallized intelligence, but rather to point out certain methodological weaknesses in this study that vitiate interesting hypothesis tests and to suggest how stronger designs could be obtained.

Central to the logic of intervention research is the establishment of just what it is that is modified. The behavioral data of intelligence are typically scores on one or more abilities measures. There are factors in intervention studies that militate against the automatic assumption that improvement on test performance reflects improvement on the ability assessed by the test. Such factors include an increasing familiarity of test stimuli with practice and a general increase in test-manship. Particularly vexatious is the question of whether an observed effect is a genuine intellectual increment or the result of learning a number of rather specific tricks that are useful in solving problems of a certain type — in other words, coaching. To conclude that an intellectual ability has increased, it is necessary to demonstrate an improvement of sufficient breadth to exclude the rival hypothesis that only test performance on homogeneous items has increased. This requirement can be met by including enough tests to define a primary ability (French et al., 1963) and demonstrating generalized improvement on these (or on the corresponding factor score). If a broad ability, such as fluid intelligence (Gf) or crystallized intelligence (Gc), is studied, then enough tests must be included to define this second-order factor. If an ability is claimed to improve with treatment, then the breadth of the treatment effect must demonstrate that something having the structure of an ability has in fact been modified.

There is little in the Plemons et al. study to suggest that anything as broad as an ability has been modified; there is little to suggest that anything more than a practice or coaching effect has occurred. A treatment effect was demonstrated on the Figural Relations Diagnostic Test (FRDT), but not on the Cattell-Horn Tests of Figural Relations (although post-hoc analyses, in the absence of a significant condition main effect did indicate a significant difference at post-test 1). Although the Cattell-Horn battery is listed as one degree removed from the FRDT on the dimension of similarity, the extent of this dissimilarity cannot be great, since items on the FRDT were constructed to be of the same type, and to use the same relational rules, as items of the Cattell-Horn battery. Since the two measures correlated almost to the extent of their reliabilities, they could almost be considered as parallel forms of the same scale. Given that each treatment subject received 8 hours of training on items closely resembling those constituting the Cattell-Horn battery, it is surprising that the treatment effect is so hard to detect.

The size of the "treatment" effect is particularly unimpressive relative to the magnitude of the general retest effect. Performance on the Cattell-Horn battery actually improved in the control group relative to the treatment group, precisely the opposite of what would be expected if training had generalized. By the third Cattell-Horn posttest, control group and training group performances were equal. What is the point of under-

going 8 hours of training if the same level of performance can be achieved with no training? If the prima facie evidence of improved test performance is to be taken as improvement in intellectual ability, as the authors do with respect to this narrow "treatment" effect, then one would have to conclude that the most generally effective and economical way to boost intellectual ability is simply to administer the same test to the same individuals three times. It is unlikely that a modification in the primary ability of figural relations has been achieved; a more parsimonious explanation is that the training effect consists of stimuli-specific tricks and algorithms for manipulating certain kinds of figural relations problems with greater efficiency. It is clear, moreover, that nothing so broad as Gf has been modified, since performance on Induction, one of the strongest Gf markers, was completely unaffected by training. To this extent, the title and conclusions about the modifiability of Gf are misleading.

The authors relegate to a footnote a related matter of considerable importance: should normal Gf markers, such as the figural relations tests of this study, still be so considered after subjects have received substantial instruction on methods and strategies of solutions? Since these instructions must be considered primarily as cultural and educational influences, which determine Gc, there is some question whether performance on these tests after training would be characteristic of Gf, an intellectual capacity for dealing with novel or relatively culture-free material. Since the factorial structure of the variables was not reported in this study (as might be expected with groups of only 15 subjects), there is little basis for knowing whether the figural relations tests still represent an aspect of Gf.

The authors' consideration of this question is not entirely satisfactory. They state that the correlations between the Cattell-Horn battery and Verbal Comprehension (a measure of Gc) did not differ for the treatment and control groups, but that the correlation between Induction (a measure of Gf) and the Cattell-Horn battery for the treatment group ($r = .74$) exceeded that for the control group ($r = .37$). These results were said to support a claim that the Cattell-Horn battery and the similar FRDT could still be considered Gf measures, even though the correlations of Cattell-Horn test of Figural Relations with Induction are *not* significantly different in the two groups (using Fisher's transformation, $p = .18$, two tailed). With sample sizes this small ($n_1 = n_2 = 15$), and without other marker variables for comparison, one simply has no way of knowing whether the figural relations tests for the treatment group would load on Gf or not.

The authors argue that their results are relevant to theories, such as that of fluid and crystallized intelligence, that specify normative age trends with respect to a particular ability or trait. Demonstrated improvement of Gf tasks in aged adults is apparently thought to be inconsistent with the normative age decline observed for this ability. Grant, for the sake of argument, that Gf, rather than test performance, had been improved. This would certainly refute a theory of immutable decline. But the assumption that age trends of variables possessing traitlike properties must be immutable, or resistant to modification, or even resistant to short-term modification, is false. Visual acuity, for example, declines with age in the population, yet this is no less true when it is realized that immediate beneficial modifications in this ability can accrue to many individuals by a proper selection of eyeglasses. The frequently observed age decline in Gf tasks is probabilistic (i.e., true on the average), not immutable (e.g., Botwinick, 1977; Horn & Donaldson, 1976, 1977). Horn (e.g., 1970, 1975) has consistently argued that some individuals may decline less in Gf than others, and that some may avoid Gf decline entirely. Intervention studies conceived as tests of such normative age trends are thus susceptible to criticisms concerning straw man architecture.

Furthermore, the evidence adduced by Plemons et al. does not even support the weaker claim that the Gf age trend in a hypothetical treatment population avoids the normative decline observed in the general population. Without the inclusion of younger controls in the design, there is no basis for concluding that improvement in older participants would not have been matched by comparable improvement in younger persons had they received the same training, and hence no basis for inferences about age trends in the treatment population. The control group that was included in the experiment did not (as far as one can tell from the authors' description) receive as much experimental attention, or evidence of personal concern, as the treatment group. Treatment effects are therefore confounded with such differences: improvement in performance of the treatment group could have been the result of their greater motivation to do well for sympathetic experimenters who were obviously trying to help them. This unwanted systematic effect could be eliminated by meeting with the controls in discussion groups, or a similar forum, for comparable periods of time.

In the absence of proper controls and an adequately defined ability factor, the title of the study, "Modifiability of fluid intelligence in aging," cannot be taken seriously. To test the effects of training on abilities, as these are understood in the usual factor-analytic sense, information about the factor structure in both treatment and control groups is needed. This information can be obtained only if enough tests are included to define the abilities adequately, and if enough individuals are measured so that there is some confidence in the stability of the correlations and factor leadings. Under these conditions, hypotheses derived from Gf-Gc theory (for example) could be clearly formulated and tested. It would not be necessary to speculate about whether improvement was manifested in something as broad as a factorial ability, or about whether training had altered the pattern of factor loadings of a test. If the factor structure is not changed by the treatment, then a comparison of factor scores for treatment and control groups would permit inferences about change in ability to be drawn with some confidence; if the factor structure is changed by treatment then the nature of the change is itself interesting. Although these analyses are not without problems, the problems are no more serious when made explicit (see, for example, Rock et al., 1978) than when obscured in experiments that offer no basis for consideration of the interesting hypotheses discussed above. Further improvements in design could be effected by including younger controls, which would allow inferences to be drawn about normative age trends in the hypothetical treatment population, and by controlling for noncognitive influences (such as differences in motivation) associated with unequal experimenter contact with treatment and control groups.

These suggestions are neither definitive nor exhaustive; other researchers would probably offer additional suggestions of equal or greater value (the importance of temporal stability, for example, has correctly been stressed by Plemons et al.). They are offered as constructive guidelines for future research in this area. It may not always be practical to include all these features in any one experiment, but it is hoped that consideration of the suggested improvements will lead to designs of increased sophistication, which may reveal valuable information about how intellectual abilities can be modified.

Gary Donaldson, PhD
Univ. of Washington

REFERENCES

Botwinick, J. Aging and intelligence. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging*. Van Nostrand Reinhold, New York, 1977.

French, J. W., Ekstrom, R. B., & Price, L. A. *Manual and kit for cognitive factors*. Educational Testing Service, Princeton, NJ, 1963.

Horn, J. L. Organization of data on life-span development of human abilities. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology*. Academic Press, New York, 1970.

Horn, J. L. Psychometric studies of aging and intelligence. In S. Gershon & A. Raskin (Eds.), *Aging, Volume 2: Genesis and treatment of psychologic disorders in the elderly*. Raven, New York, 1975.

Horn, J. L., & Donaldson, G. On the myth of intellectual decline in adulthood. *American Psychologist*, 1976, *31*, 701-719.

Horn, J. L., & Donaldson, G. Faith is not enough: A response to the Baltes-Schaie claim that intelligence does not wane. *American Psychologist*, 1977, *32*, 369-373.

Plemons, J. K., Willis, S. L., & Baltes, P. B. Modifiability of fluid intelligence in aging: A short-term longitudinal training approach. *Journal of Gerontology*, 1978, *33*, 224-231.

Rock, D. A., Werts, C. E., & Flaugher, R. L. The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, 1978, *13*, 403-418.

Dear Editor:

We welcome this opportunity to reply to Donaldson's critique. The Plemons, Willis, and Baltes article (1978) reported on results of a pilot study initiating an extensive 5-year program of cognitive training research conducted to examine the modifiability of older adults' performance on several fluid intelligence dimensions. This research program has involved a series of studies, including a replication/extension of the Plemons et al. study (Willis et al., 1981), training research on fluid-related dimensions of inductive reasoning (Blieszner et al., in press) and attentional processes (Baltes & Willis, 1981), examination of retest/practice effects on fluid dimensions (Hofland et al., 1981), and structural analyses of Gf-Gc ability relationships in later life (Baltes et al., 1980). The Donaldson paper raises several critical issues for training research, but we differ with many of his conclusions. We find the timing of this exchange opportune for examining these issues with regard to the Plemons et al. study and our subsequent research.

Donaldson focuses on four major issues: (1) breadth of transfer effects; (2) assessing training effects as a function of ability-related factors vs. ability-extraneous variables, such as test sophistication and motivation; (3) possible changes in measurement validity related to training; (4) normative age trends and the appropriate use of younger control groups.

He first questions whether training improvement should be interpreted as representing ability-specific effects. He suggests that training gains may be largely attributable to ability-extraneous factors, such as test sophistication or increased motivation, that affect test performance but are not specific to the figural relations ability per se. We have examined training effects using a set of three criteria: (a) breadth of transfer within the target ability, (b) a hierarchical pattern of training transfer across fluid-crystallized intelligence (Gf-Gc) dimensions; (c) temporal maintenance of training effects. We believe consideration of all three criteria provides a more systematic and exacting assessment paradigm than Donaldson's more limited focus on only measurement breadth.

As to the issue of *breadth* of transfer emphasized by Donaldson, our focus was on the range of transfer *within* the figural relations ability. We hypothesized that training effects would be evident for the two measures of the figural relations ability: the Figural Relations Diagnostic (FRDT) designed by us and the Cattell-Horn measures of figural relations originally developed by Cattell and Cattell (1961) and known as the Culture Fair Test. Training effects were found on the FRDT on all three posttest occasions and on the first posttest for the Cattell-Horn battery. Donaldson was not impressed with this effect, we believe, in part because he is not sufficiently familiar with the tasks and tests involved. Contrary to what Donaldson suggests, content validity of these two tests (FRDT, Cattell-Horn battery) does not indicate that they are quasi-parallel forms of the same measure. The FRDT involves items based on the same relational rules used in designing the training items. Note, however, that none of the items are the same for training and assessment — only that the items were based on the same relational rules. In contrast to the FRDT, the Cattell-Horn battery included items involving relational rules *not* used in training or on the FRDT. Also, the Cattell-Horn battery included a subtest (Power Matrices) taken from a more advanced and more difficult scale (Scale 3) of the Culture Fair test. Performance on the Cattell-Horn battery, therefore, required participants to solve items involving relational rules never taught and to work more advanced problems than those included in training. The issue of breadth of transfer is further supported by findings from our recent replication/extension study (Willis et al., 1981). In this study, the assessment battery was broadened to examine near transfer effects across three measures of figural relations: the FRDT and Cattell-Horn battery employed in Plemons et al. and, in addition, the Raven's Advanced Progressive Matrices. The Raven's involved both items with different relational rules, and also much more difficult items than contained in the other two tests. Significant training effects were found on all three of these measures and

were maintained at the six-month posttest. Note that the number of tests used to assess training is equivalent to or greater than the number of measures employed to mark an ability factor in most prior Gf-Gc research. As to magnitude of training improvement, training effects at first posttest were on the order of .9 of a standard deviation for the FRDT and approximately .5 of a standard deviation for the Cattell-Horn battery and Raven's. Such training improvement is in contrast to a retest gain of approximately .2 of a standard deviation for control. Similar patterns of breadth of transfer within the target ability have been found in our training studies on Induction (Blieszner et al., in press) and attention (Baltes & Willis, 1981).

Furthermore, we predicted that the pattern of training transfer would differ for effects associated with ability-specific vs. ability-extraneous factors. The posttest assessment included a battery of Gf and Gc measures. If training effects were ability-specific, the pattern of transfer should be *hierarchically* ordered with strongest effects for measures of Figural Relations, less transfer to the fluid ability of Induction, and no transfer to a crystallized measure of Verbal Comprehension. Such a hierarchical transfer pattern was derived from the structural pattern of ability relationships postulated by Gf-Gc theory and documented in prior factor analytic research (Cattell, 1971; Horn, 1978). Training effects were shown to be strongest for the several measures of the target Figural Relations ability. No training transfer to Gc was reported, as predicted. Therefore, hierarchical transfer involved both positive transfer to figural relations dimensions and no transfer to a Gc dimension. In contrast, if training focused largely on ability-extraneous factors, as suggested by Donaldson, then a broader transfer pattern was predicted. Since ability-extraneous factors, such as testmanship should affect performance on all or most measures, transfer associated with such factors should occur for all or most of the Gf *and* Gc measures, rather than being specific to figural relations. In line with our predictions, a hierarchical pattern of training transfer, supporting an ability-specific interpretation, was found in the Plemons et al. study. This finding was replicated in a subsequent figural relations training study (Willis et al., 1981). In contrast to the hierarchical transfer pattern for the training group, the control group, participating in only pre- and posttesting, demonstrated generalized, non-hierarchical retest effects across most measures.

Like Donaldson, we, too, were concerned about the ability-extraneous factor of motivation. Donaldson suggested that the greater social contact experienced by subjects during training sessions may have contributed to increased motivation, and, hence, improved posttest performance. We examined this issue in a recent training study focusing on attention-memory (Baltes & Willis, 1981), but the findings do not support Donaldson's hypothesis. A social contact control group received the same number of contact hours as did the training group. Significant training effects were found, but there were no significant group differences between social and no-contact controls for most measures.

Our argument for an ability-specific interpretation of the results is further supported by findings from several studies (e.g., Denney, 1980; Hoyer et al., 1973) that have trained solely on noncognitive factors (e.g., speed, motivation) assumed to affect performance on intellectual tasks. While performance on noncognitive variables has been modified, none of these studies (as reviewed by Willis & Schaie, 1981) reported significant transfer effects to individual intellectual measures.

Donaldson notes only in passing our third assessment criterion of temporal maintenance of training effects. In both the Plemons et al. and our recent replication study (Willis et al., 1981), training improvement on figural relations measures was maintained across a 6-month period. If training is said to result in reliable change in level of performance on measures of the target ability, then temporal durability of effects is critical. Such a stringent 6-month test of training maintenance has been applied or met only rarely in gerontological cognitive intervention research (see, however, Sanders & Sanders, 1978, for an exception).

There is a logical inconsistency in Donaldson's suggestion that the educational nature of the training program may have resulted in changes in the measurement validity of the figural relations tests such that at posttest they are more representative of Gc. On the one hand, Donaldson dismisses training improvement as too narrow in scope to reflect change at the level of the figural relations ability. On the other hand, he suggests that training may result in structural change in figural relations measures such that they are now more crystallized in nature. Comparisons of the Gf-Gc correlation matrices at posttest separately for training and control are in the opposite direction to Donaldson's hypotheses. First, Posttest 1 correlations between the measures of figural relations and Verbal Comprehension (Gc) did not differ by group, contrary to Donaldson's suggestion. Second, the correlation between figural relations and Induction (another Gf ability) measures was higher for the experimental group (r = .74) than for the control (r = .37). We have further examined systematically the issue of change in Gf measurement validity in our research on practice effects (Hofland et al., 1981) and in our analyses of Gf-Gc structural relationships in older adults (Baltes et al., 1980) and have found no evidence for Donaldson's hypothesis. More importantly, we believe that if it were possible to alter measurement validity through short-term training, credibility of the structural properties of the Gf-Gc theory would be brought into serious question. If Donaldson's suggestion is correct, then it would become critical for Gf-Gc theory to specify the conditions under which such validity changes could occur.

A final issue deals with the relevance of training studies for examining normative age trends. First, Donaldson fails to recognize that the major focus of the Plemons et al. study was on intraindividual variability *within* aged samples, not on normative age trends. Cross-sectional research (as usually cited by Horn and colleagues) gives information on normative interindividual differences and assumed normative intraindividual change when people are observed in the "natural" environment. This does not specify the range of intraindividual plasticity possible under more facilitative conditions. Intervention research is one method for examining the full range of possible behavior in later life. It is only when the full range of such potential behavior is examined that so-callaed normative age trends can be put into proper perspective. We believe that the primary emphasis in prior gerontological literature on describing normative patterns of decline has hindered an interventive, facilitative approach

to intellectual aging and has led the layman to assume such decrement is irreversible.

Donaldson's reasoning on the use of younger controls to examine age trends has serious flaws. The key question in the use of younger controls in aging research is to examine what such groups can assess or control for (Baltes et al., 1977). Use of younger controls can be seriously misleading in studying the range of modifiability of behavior in current older cohorts. Age differences based on cross-sectional data are indicative of a large class of life history differences, not only ability differences. Further, it is unclear whether current elderly cohorts (Schaie, 1979) performed at comparable intellectual levels with today's youth when at a younger age, as Donaldson's argument would require. Assuming, as Donaldson does, that a younger control group would provide information on "true" age change in ability is naive and reflects a misunderstanding of the complexities involved in developmental research.

In this reply, we have discussed four major issues raised by Donaldson regarding training assessment, breadth of transfer, measurement validity, and use of younger controls. It has been shown that a paradigm for examining training assessment was defined more carefully in the Plemons et al. study than Donaldson's critique implies or than his own suggestions for improvement would require. Further, considerable breadth of transfer was demonstrated if consideration is given to the content validity of the figural relations measures utilized in Plemons et al. and findings from our replication study. Finally, we find Donaldson's proposals regarding change in measurement validity and use of younger controls to have serious theoretical and/or methodological flaws. In summary, we reject most of Donaldson's conclusions. We reaffirm our belief in the importance of intervention research as one methodological procedure for examining the full range of behavior in older adults and strongly urge further research of this type.

Sherry L. Willis, PhD
The Pennsylvania State Univ.

Paul B. Baltes, PhD
Max Planck Inst. for Human
Dev. & Educ.

REFERENCES

Baltes, P. B., Cornelius, S. W., Nesselroade, J. R., & Willis, S. L. Integration vs. differentiation of fluid-crystallized intelligence in old age. *Developmental Psychology*, 1980, *16*, 625-635.

Baltes, P. B., Reese, H. W., & Nesselroade, J. R. *Life-span developmental psychology: Introduction to research methods*. Brooks Cole, Monterey, CA, 1977.

Baltes, P. B., & Willis, S. L. Enhancement (plasticity) of intellectual functioning in old age: Penn State's Adult Development & Enrichment Project (ADEPT). In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes*. Plenum, New York, 1981. (in press)

Blieszner, R., Willis, S., & Baltes, P. Training research in aging on the fluid ability of inductive reasoning. *Journal of Applied Developmental Psychology*, 1981. (in press)

Cattell, R. B. *Abilities: Structure, growth, and action*. Houghton-Mifflin, New York, 1971.

Cattell, R. B., & Cattell, A. K. Measuring intelligence with the Culture Fair Test. *Manual for Scales 2 and 3*. IPAT, Champaign, IL, 1961.

Denney, N. W. The effect of manipulation of peripheral, noncognitive variables on the problem-solving performance of the elderly. *Human Development*, 1981, *23*, 268-277.

Hofland, B. F., Willis, S. L., & Baltes, P. B. Fluid intelligence performance in the elderly: Retesting and intraindividual variability. *Journal of Educational Psychology*, 1981. (in press)

Horn, J. L. Human ability systems. In P. B. Baltes (Ed.), *Life-span development and behavior*, Vol. 1, Academic Press, New York, 1978.

Hoyer, W. F., Labouvie, G. V., & Baltes, P. B. Modification of response speed deficits and intellectual performance in the elderly. *Human Development*, 1973, *16*, 233-242.

Plemons, J. K., Willis, S. L., & Baltes, P. B. Modifiability of fluid intelligence in aging: A short-term longitudinal training approach. *Journal of Gerontology*, 1978, *33*, 224-231.

Sanders, R. E., & Sanders, J. C. Long-term durability and transfer of enhanced conceptual performance in the elderly. *Journal of Gerontology*, 1978, *33*, 408-412.

Schaie, K. W. The primary mental abilities in adulthood: An exploration in the development of psychometric intelligence. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior*, Vol. 2, Academic Press, New York, 1979.

Willis, S. L., Blieszner, R., & Baltes, P. B. Intellectual training research in aging: Modification of performance on the fluid ability of figural relations. *Journal of Educational Psychology*, 1981, *73*, 41-50.

Willis, S. L., & Schaie, K. W. Maintenance and decline of adult mental abilities: II. Susceptibility to experimental manipulation. In F. Grote (Ed.), *Ninth symposium on learning: Adult learning and development*. Western Washington Univ., Bellingham, WA, 1981. (in press)