Universität Potsdam

Paul B. Baltes, Doris Sowarka, Reinhold Kliegl

Cognitive training research on fluid intelligence in old age : what can older adults achieve by themselves?

Cognitive Training Research on Fluid Intelligence in Old Age: What Can Older Adults Achieve by Themselves?

Paul B. Baltes, Doris Sowarka, and Reinhold Kliegl
Max Planck Institute for Human Development and Education, Berlin, Federal Republic of Germany

Cognitive research on the plasticity of fluid intelligence has demonstrated that older adults benefit markedly from guided practice in cognitive skills and problem-solving strategies. We examined to what degree older adults are capable by themselves of achieving similar practice gains, focusing on the fluid ability of figural relations. A sample of 72 healthy older adults was assigned randomly to three conditions: control, tutor-guided training, self-guided training. Training time and training materials were held constant for the two training conditions. Posttraining performances were analyzed using a transfer of training paradigm in terms of three indicators: correct responses, accuracy, and level of item difficulty. The training programs were effective and produced a significant but narrow band of within-ability transfer. However, there was no difference between the two training groups. Older adults were shown to be capable of producing gains by themselves that were comparable to those obtained following tutor-guided training in the nature of test-relevant cognitive skills.

Several research programs have demonstrated that older adults have a substantial reserve capacity and are capable of increasing their performance on tests of fluid intelligence as a consequence of guided instruction in problem-solving strategies by an "expert" tutor (Baltes & Lindenberger, 1988; Willis, 1987). What is not known, however, is to what degree older adults are able to generate similar training benefits by themselves without guided instruction in test-relevant cognitive skills by a tutor. Research on memory functioning (Kausler, 1982; Treat, Poon, & Fozard, 1981), for example, has demonstrated that some older adults are able to generate useful learning strategies on their own.

As to the specific domain of this study, there are three lines of evidence that suggest that older adults may have a sizeable reserve capacity for generating higher levels of test performance by themselves without much external guidance in the nature of problem solving. The first involves positive retest effects associated with repeated administration of the same tests of fluid intelligence (Hofland, Willis, & Baltes, 1981). This work, however, is restricted to test-specific gains as it typically does not consider transfer to a new set of items. The second line of evidence involves the use of power conditions of assessment such as extending the time available for solving tests of intelligence. When older adults were given more time for item solution (such as a fourfold increase in testing time), they displayed increments in performance of a magnitude that were similar to the training gains obtained following guided instruction in test-relevant cognitive skills (Baltes, Kliegl, & Dittmann-Kohli, 1988). The third line of research involves the use of training programs of fluid intelligence varying widely in content and format. These programs include training in test-relevant cognitive skills, reduction of test anxiety, and the use of group-based discussion about appropriate problem-solving strategies (Blackburn, Papalia-Finlay, Foye, & Serlin, 1988; Hayslip, 1989; Willis, 1987). In this line of work, each of these rather diverse training programs resulted in significant gains in test performance. Such a pattern of results seems plausible if older adults possess the cognitive repertoire or reserve capacity necessary for higher performance to begin with.

The purpose of this study was to examine further the extent to which older adults by themselves possess the cognitive reserve capacity necessary to achieve gains in tests of fluid intelligence. In order to obtain such information, a new training program was developed whose primary focus was self-guided practice. In this new training program, older adults were given the opportunity to practice criterion-relevant materials, without guidance by a tutor, in the nature of problem solving. To assess the compositional nature of posttraining performance and to obtain a fine-grained assessment of training benefits, we examined several indicators of training gain using a transfer of training paradigm. In past research (Baltes, Dittmann-Kohli, & Kliegl, 1986; Baltes et al., 1988), it has been shown that, in addition to items correct, it is useful to consider error patterns and the level of difficulty of the test items solved. Therefore, the following indicators of performance after training were assessed: (a) amount of transfer to a new set of test items, (b) accuracy of performance, and (c) level of difficulty of the items solved correctly.

Method

The present study involved three experimental groups (two training, one control) and a pretest–posttest arrangement with random assignment of subjects to experimental groups following the pretest.

Subjects

The study sample of city-dwelling (West Berlin) elderly adults was recruited by newspaper advertisements in the same manner as earlier research on the plasticity of fluid intelligence in old age (Baltes et al., 1986). All of the subjects were "naive" volunteers and were physically able to come to the laboratory. They were paid an amount of DM150 (approximately $80) as a fee and to cover their travel expenses. When compared with the general population of elderly people living in the city, the present sample was positively biased in terms of health and education. Such a positive bias was desired not only to achieve comparability with earlier research but also to minimize the likelihood of having a sample with a high proportion of disease-related (secondary) aging.

The final sample ($N = 72$) consisted of 52 female and 20 male older adults (M age = 72 years; range = 63–90 years). Their reported subjective health (on a 5-point scale) was above average ($M = 3.6$, $SD = 0.7$). Average educational level (roughly comparable with U.S. information on educational history) was 10.7 years ($SD = 2.0$). These indicators closely parallel those obtained in earlier research conducted in the present laboratory on the same topic.

The final sample was based on an initial contact sample of 101 volunteering subjects who appeared for the first test session in the laboratory. To minimize the possible effect of selective dropout (Baltes, Reese, & Nesselroade, 1977), random assignment occurred after completion of a two-session pretest. Of the 101 subjects volunteering, 94 completed the second session of the pretest. After this second pretest, subjects were randomly assigned to three experimental conditions. Of the 94 subjects volunteering for the remainder of the study, 77 subjects completed all of the sessions. Five subjects chosen at random were dropped to achieve three samples with equal sizes of $n = 24$ each.

Dropout analyses were performed to assess whether dropout was selective using three indicators: age, subjective health, and education. Analyses of variance revealed no difference in the respective means of the initial dropouts and the sample completing the study (age = 72.4 vs. 71.8 years; health = 3.6 vs. 3.7; education = 10.7 vs. 10.6 years). Also, no differences were obtained when comparing the dropouts from the next-to-final sample ($n = 17$) with the subjects ($n = 77$) who had complete data protocols. This latter comparison included a comparison of IQ (as estimated by a standardized mean score across all subtests of intelligence).

Procedure and Design

The experimental–control-group design involved three main parts: a pretest, cognitive training (two training conditions vs. a no-contact control), and a posttest administered 1 week after training. The test battery given at pre- and posttest consisted of six subtests, which will be described later. The no-contact group received no treatment intervening between pre- and posttest. Earlier research (Blieszner, Willis, & Baltes, 1981) had shown that a contact-control group was not necessary.

Measures

At pretest and posttest, a battery of tests of intelligence (3 hr, 45 min) was administered in two sessions of about equal length. Following the rationale presented in earlier publications (e.g., Baltes et al., 1986; Baltes & Willis, 1982), we chose the battery to measure the domain of fluid and crystallized intelligence so that a fine-grained analysis on the level of subabilities of fluid intelligence—that is, the domain that represented the focus of training—was possible.

The test battery included all of the measures used in past research (Baltes et al., 1986) except for the measures of perceptual speed. In past research, perceptual speed measures were shown not to be affected by the kind of cognitive training administered in this study (Baltes & Lindenberger, 1988). Depending on their similarity to the training program, the six tests of the battery can be ordered along a continuum of transfer from near to far transfer (Baltes & Willis, 1982; see also the order of tests in Figure 1).

Cognitive Training Programs

The training subjects participated in one of two cognitive training programs, each consisting of five training sessions of about 1 hr in length. Both training programs focused on one of the two major subabilities defining fluid intelligence, namely, figural relations.

The two figural relations training programs used the same practice materials (test items) and the same training schedule (five sessions of about equal length). They differed in the extent to which instruction was given on the nature of effective problem solving and feedback about the correctness of the problems practiced during training.

Tutor-guided training in figural relations. The tutor-guided program consisted of the delivery of a training module developed earlier and effectively used in past research (Baltes & Willis, 1982; Willis, Blieszner, & Baltes, 1981). Its emphasis was on the teaching of a set of test-relevant problem-solving skills and feedback about the correctness of practiced test solutions. This figural relations training program was developed through task analysis of the items and rules contained in established tests of figural relations. None of the items used in training, however, was identical in content to the items constituting the standardized test forms used to assess training effectiveness.

The training program involved five sessions. The first four sessions dealt with one subcomponent of the test domain of figural relations. The fifth session consisted of a review of all materials, rules, and concepts. The training program, conducted in small groups of 5 to 10 participants, focused on helping subjects to identify rules and concepts and to use them in solving the types of problems associated with tests of figural relations. The tutor and participating subjects modeled how to identify and use the rules; subjects practiced using these strategies. Feedback and discussion about these strategies, as well as about the possibility of using alternate strategies, followed. Further information about the training program is contained in Willis et al. (1981).

Self-guided practice. The self-guided training program was developed for the present study. It used the same practice materials (problems) as the tutor-guided training module. Self-guided practice sessions were somewhat shorter (5–15 min) than the tutor-guided sessions. This was due to the length of time the tutor needed in the tutor-guided program when explaining problem-solving strategies and giving feedback on individual problem solutions.

The self-guided training group received a minimum of instruction about the nature of problem solving and the skills involved. The tutor restricted her comments to introducing the problems by use of two to three warm-up items to ensure that the format of test items was understood (using the same warm-up items that preceded the standard form of the tests involved). Subsequently, the tutor indicated that the subjects (physically well separated from each other) were to work individually on the practice booklets without outside assistance. It was stated explicitly and repeatedly that the purpose of the study was to examine how well subjects could do on their own. No feedback about the nature of problem solving and the correctness of problem solutions was given.[1]

Communication among the subjects was restricted to matters of daily life and social conversation. On the basis of observations by a proctor (one of the authors) and private interviews with a convenient subsample of the training group, we have no reason to believe that the subjects departed from the intent of this instruction.

Results

Data Analyses

To provide comparability with earlier studies, we transformed the data into a common metric for all tests ($M = 50$,

---

[1] A manual describing the training program is available from the authors.

SD = 10), using the performance of the control group at pretest in Baltes et al. (1986) as baseline. This approach permitted, first, a direct comparison between the level of performance displayed by the present sample at pretest and that of earlier work conducted in our laboratory as a measure of sample comparability. In the present instance, the overall level of performance of the study sample at pretest was almost identical ($M = 49.56$, $SD = 9.8$) with the Baltes et al. (1986) control group used as a standard of comparison. Second, the use of the standardization procedure also permitted a direct comparison of magnitude of training gains with the gains obtained in past work.

Because of the nature of our predictions and desired comparability with past work, we tested the hypotheses by means of a priori specified contrasts. A first set of contrasts involved the three groups in the experiment. Two contrasts were specified, one testing the difference between the two cognitive training groups, another testing the difference between these two groups and the no-contact control group. A second set of three contrasts concerned the nature of transfer along the postulated continuum of transfer tests. Following the logic of earlier studies and the present design (see Figure 1 for order of tests from 1 through 6), a first contrast tested all levels of fluid transfer against nonfluid transfer. This contrast involved the first five tests against the sixth. A second contrast tested two levels within near-fluid transfer against each other, that is, Tests 1 and 2 against Tests 3 to 5. In this contrast, we excluded the Raven from the nearest category of transfer, because in our laboratory the Raven was never affected by training. The third contrast concerned two levels within near-near transfer (Test 1 against Test 2).

The third set of contrasts involved level of difficulty of the items solved during pretest and posttest. Two contrasts were specified to disentangle differential effects of training: easy and medium against difficult, and easy against medium.

Training Effects and Pattern of Transfer

Of initial concern to the present study was whether application of the standard module of tutor-guided training of figural relations produced results similar to past work. Subsequently, the central question was whether the training effects obtained showed differences between the two training groups. Figure 1 displays the profile of gains on the six tests for the three groups. Training was effective for the two near-transfer measures. The two training groups did not differ from each other. This conclusion is based on the following statistical analyses.

As was true for earlier work, change (or difference) scores from pretest to posttest were used as dependent variables.[2] All three contrasts involving the ordering of tests along the transfer continuum (1–6) were significant. Improvement on all fluid transfer tests (1–5) tended to be larger than those on the nonfluid vocabulary test, $F(1, 69) = 3.6$, $p < .06$. Gains were also larger on figural relations tests (1 and 2) than on the other fluid transfer tests (3–5), $F(1, 69) = 29.4$, $p < .01$. Finally, improvement was larger on the nearest transfer test, the direct focus of training (ADEPT [Adult Development and Enrichment Project] figural relations; Baltes & Willis, 1982) than on the second figural relations test (Culture Fair Test [CFT]), $F(1, 69) = 8.5$, $p < .01$.

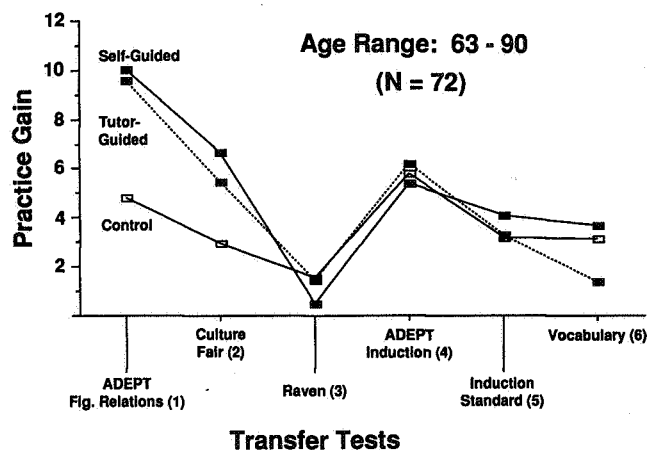What about the contrasts involving the three experimental

Figure 1. Pattern of transfer (1–6) of training gains (correct responses) separately for two training groups (tutor-guided, self-guided) and a control group.

groups? Using the criterion of a 5% level of significance, none of the main effects of group were significant. The results for the contrast between the two training groups versus the control group were $F(1, 69) = 3.4$, $p > .08$, and the results for the contrast between the two training groups were $F(1, 69) = 0.4$, $p > .5$. As expected, there was one significant interaction contrast. Both training groups profited more than the control group on the two transfer tests (ADEPT figural relations and Culture Fair) that were the focus of training relative to the performance changes associated with other fluid transfer tests (3–5), $F(1, 69) = 8.8$, $p < .01$. None of the interactions tested, involving the tutor-guided and the self-guided training groups, were significant (all $Fs < 1$).

This first set of analyses showed that, when examining levels of performance between the two training groups using the number of correct answers as dependent variables, both training groups exhibited significant gains on the two nearest within-ability transfer measures. In addition, the analyses demonstrated that the two training groups performed at the same level.

What about the magnitude of the training gains achieved for the training groups and its similarity to past work? The magnitude of training gains observed on the nearest transfer measure is about one half of a standard deviation expressed in the units of the standardization taken from Baltes et al. (1986). This amount is comparable with training gains achieved in past work using the same tutor-guided training module in figural relations. Thus, the absence of a difference between the two training groups does not seem to reflect the fact that the tutor-guided training program was less effective than in the past.

Accuracy of Performance

Did training participants tutored in test-relevant skills commit fewer false responses relative to their increase in correct responses than subjects participating in self-guided training? To

---

[2] When statistical analyses were performed with analysis of covariance (using the pretest scores as covariates), the same outcomes resulted.

examine this question we computed a dependent variable reflecting accuracy. Accuracy scores were computed as the number of correct answers divided by the sum of the correct and false answers (errors of commission).

Of course, such accuracy scores are not independent from the dependent variable in the previous analysis. Yet, given that the two training groups had identical profiles regarding the number of correct responses, differences between the groups in accuracy scores were likely due to differences in the amount of error. The statistical analysis performed on the accuracy scores yielded no difference (all $Fs < 2.1, p > .10$).[3]

Item Difficulty

The remaining indicator of performance concerned the level of difficulty of test items. From an earlier study (Baltes et al., 1986), estimates of item difficulty were available for three of the near-transfer tests (ADEPT figural relations, Culture Fair, and ADEPT induction). Items making up each of the three tests were categorized into thirds (easy, medium, and difficult items) on the basis of their probability of being solved by a comparable elderly sample under a power (time-extended) condition of assessment. On the basis of this classification, scores were computed separately for easy, medium, and difficult items.

Pre- to posttest gains were significantly larger in items of easy and medium difficulty when compared with items of high difficulty: $F(1, 69) = 16.2, p < .01$. In addition, pre- to posttest gains were significantly larger in items of medium level of difficulty when compared with the easiest category, $F(1, 69) = 6.2, p < .02$. Of the interactions involving group, only one was significant, showing that both training groups improved more on items of easy and medium difficulty than did the control group, $F(1, 69) = 8.3, p < .01$. The three groups did not differ in their performance on the most difficult items. (When these analyses were performed on accuracy scores, the same outcomes resulted.)

In general, then, the performance gains for the two training groups were identical. Furthermore, the locus of performance gains was in the categories of easy and medium level of item difficulty. Findings based on a post hoc inspection of all data matrices suggested one departure from this pattern. Tutor-guided participants seemed to solve more items of medium difficulty than did the self-guided group on one of the three tests, namely, the ADEPT figural relations. The relevant triple interaction of group, test, and difficulty had not been significant, $F(1, 69) = 0.70, p > .40$. However, a direct post hoc comparison revealed a significant difference in favor of the tutor-guided group, $t(46) = 2.5, p < .02$. We tend to interpret this finding as due to chance.

Power Condition of Assessment

The present experiment involved an additional condition. At the conclusion of the experiment as reported here, subjects also answered the near-transfer tests under a quadruple-power condition (four times the usual testing time). The purpose of this power condition of assessment (Baltes, Kliegl, & Dittmann-Kohli, 1988; Willis, 1987) was to examine whether training gains were due primarily to an increase in the speed of responding and the ensuing possibility of trained subjects being able to

attempt more problems within the time constraints of standardized testing.

The analysis of these data did not alter the pattern of the results reported here for time-limited assessment. In other words, even when subjects had more time available at posttraining assessment, there was no difference between the tutor-guided and the self-guided group in each of the three indicators examined (number correct, accuracy, and level of difficulty).

Conclusions and Discussion

The answers to our research questions seem clear. Within the design confines of the present study, elderly subjects were capable by themselves of producing training gains that were identical with those observed for the group that participated in a tutor-guided training program in which the explicit focus was on test-relevant cognitive skills and feedback about performance. The various analyses conducted on several indicators of performance (correct, accuracy, level of difficulty, and power condition) support, with increasing certainty, the conclusion of no difference between the two training groups.

First, there is the fact that tutor-guided practice of test-relevant cognitive skills produced a significant amount of transfer similar to what was obtained in earlier research (Baltes et al., 1986; Baltes & Willis, 1982). With this finding, an appropriate baseline for comparison with past work was successfully established. Note also that the training gains obtained in the present study were not simple retest gains based on the same test materials. Rather, gains involved transfer to a new set of test items not contained in the training materials.

Second, the interpretation that elderly adults possess the requisite cognitive skills in their repertoire and therefore are capable of producing comparable training gains by themselves, is strengthened by the results on accuracy and level of difficulty. There is no evidence that tutor-guided instruction and feedback about correctness of answers resulted in more accuracy or the ability to solve more difficult items than was true for self-guided practice. Third, tutor-guided and self-guided subjects continued to behave identically when given the opportunity to solve tests under a time-extended power condition of assessment after the conclusion of the main experiment.

Results of any empirical study have only limited generalizability. The first limiting factor of the present study concerns the fact that, when compared with the total population of their age peers, our research participants were of superior health and higher educational status. Thus, we do not know whether similar results would have been obtained had a more representative sample of elderly adults been included in the study. A second limiting factor may be the lack of several posttests in order to examine the time course of the maintenance of training effects. Because our own past research with similar training arrangements had not demonstrated an advantage of tutor-guided ability training over retest practice groups in the extent of follow-up maintenance of training gains, we had decided to restrict follow-up assessment to one immediate posttest. Meanwhile,

---

[3] When statistical analyses were performed with change scores involving false answers as sole dependent variables, the same result was obtained. There was no difference between the two training groups.

however, Blackburn et al. (1988) have reported data showing that self-generated strategies for task solution may be more durable over time than specific rule training. Thus, it is possible that a difference between the tutor- and self-guided training programs would emerge at a later posttest occasion.

The third factor limiting the generalizability of the present experiment involves the comparison between tutor-guided training and self-guided instruction. Two aspects are involved. The first is the kind of tests used, namely, standardized tests of fluid intelligence. Other cognitive tasks or tests, such as Piagetian measures of abstract logical reasoning, may have produced a different picture. The second aspect deals with the training programs themselves. The present comparison involved one subability of fluid intelligence and one type of tutor-guided training. Therefore, as we proffer a particular interpretation based on the present experiment, we need to be aware of the possibility that the use of other measures of cognitive functioning or other types of training of test-relevant cognitive skills may have resulted in a different outcome.

In conclusion, then, we certainly would not want to argue that elderly subjects are always capable of producing training gains by themselves comparable with what tutor-based practice can do. In the present context, however, the evidence points overwhelmingly in the direction of such a conclusion. It was demonstrated that elderly adults have a substantial reserve capacity. It includes the generation of training gains in transfer tests of fluid intelligence that, in the past, we had attributed primarily to the effective delivery of a tutor-based training program in test-relevant skills and external feedback about the correctness of test solutions.

References

Baltes, P. B., Dittmann-Kohli, F., & Kliegl, R. (1986). Reserve capacity of the elderly in aging-sensitive tests of fluid intelligence: Replication and extension. *Psychology and Aging, 1,* 172–177.

Baltes, P. B., Kliegl, R., & Dittmann-Kohli, F. (1988). On the locus of training gains in research on the plasticity of fluid intelligence in old age. *Journal of Educational Psychology, 80,* 392–400.

Baltes, P. B., & Lindenberger, U. (1988). On the range of cognitive plasticity in old age as a function of experience: 15 years of intervention research. *Behavior Therapy, 19,* 283–300.

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

Baltes, P. B., & Willis, S. L. (1982). Plasticity and enhancement of intellectual functioning in old age: Penn State's Adult Development and Enrichment Project (ADEPT). In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes* (pp. 353–389). New York: Plenum Press.

Blackburn, J. A., Papalia-Finlay, D., Foye, B. F., & Serlin R. C. (1988). Modifiability of figural relations performance: A combination of two training approaches. *Journal of Gerontology, 43,* 87–89.

Blieszner, R., Willis, S. L., & Baltes, P. B. (1981). Training research in aging on the fluid ability of inductive reasoning. *Journal of Applied Developmental Psychology, 2,* 247–265.

Hayslip, B., Jr. (1989). Alternative mechanisms for improvements in fluid ability performance among the aged. *Psychology and Aging, 4,* 122–124.

Hofland, B. F., Willis, S. L., & Baltes, P. B. (1981). Fluid intelligence performance in the elderly: Intraindividual variability and conditions of assessment. *Journal of Educational Psychology, 73,* 573–586.

Kausler, D. H. (1982). *Experimental psychology and human aging.* New York: Wiley.

Treat, N. J., Poon, L. W., & Fozard, J. L. (1981). Age, imagery, and practice in paired-associate learning. *Experimental Aging Research, 7,* 337–342.

Willis, S. L. (1987). Cognitive training and everyday competence. In K. W. Schaie (Ed.), *Annual review of gerontology and geriatrics* (Vol. 7, pp. 159–188). New York: Springer.

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