Assessment of remote memory: Validation of the television test by repeated testing during a 7-year period

LARRY R. SQUIRE and MARY M. FOX

Veterans Administration, San Diego, California 92161 and

Department of Psychiatry, University of California School of Medicine, La Jolla, California 92037

Remote memory for former, one-season television programs was assessed on four different occasions during 7 years. Each test yielded a forgetting curve that was primarily monotonic and linear. In addition, the forgetting curve obtained by the first application of this test in 1973 was largely recapitulated by the three later applications of the test. These findings, together with new information concerning the equivalence of items from different time periods, provide strong evidence that the television test validly measures remote memory.

Objective tests of remote memory have employed questions about past news events (Botwinick & Storandt, 1974; Squire & Cohen, in press; Warrington & Silberstein, 1970), famous faces (Albert, Butters, & Levin, 1979; Marslen-Wilson & Teuber, 1975; Sanders & Warrington, 1971), high school classmates (Bahrick, Bahrick, & Whittlinger, 1975), and former one-season television programs (Squire & Slater, 1975). Each test has different advantages and disadvantages, depending on the experimental question under study. One useful feature of the television test is that it has been constructed to satisfy the criterion of equivalence (Squire & Cohen, in press), which requires that events selected from different time periods were initially learned to a similar degree. Moreover, the television test typically yields a forgetting curve with normal subjects, such that scores are generally best for programs broadcast recently and poorest for programs broadcast long ago (Harvey & Crovitz, 1979; Squire, Chace, & Slater, 1976; Squire & Cohen, 1979; Squire & Slater, 1975; Squire, Slater, & Chace, 1975). In addition, performance on the test is positively related to the length of time a program remained on the air and to the amount of reported television viewing (Harvey & Crovitz, 1979).

Despite attempts at rigorous test construction, all remote memory tests have necessarily depended on retrospective methods. Accordingly, the television test, like other remote memory tests, has remained vulnerable to the criticism that features of the for-

getting curve could reflect idiosyncrasies of test items rather than a description of remote memory (Harvey & Crovitz, 1979; Levin, Grossman, & Kelly, 1977). Here, we report results of four successive administrations of the television test during a period of 7 years. With these repeated assessments of remote memory using updated versions of the same test, together with new information concerning the equivalence of items from different time periods, it has been possible to demonstrate validity of the test instrument and to obtain a reliable description of remote memory.

METHOD

Since the television test was first constructed in 1973 (Squire & Slater, 1975), it has been updated each year following a fixed procedure and it has been administered on three additional occasions, in 1975 (Test 2), in 1978 (Test 3), and in 1979 (Test 4). Each October, at the beginning of the new television season, a list was compiled of programs that had begun the previous October on weekly television (between 8:00 p.m. and 11:00 p.m.), but which had been canceled within 12 months. These lists excluded programs whose names were also the names of movies, popular songs, or celebrities. For the 6 years 1974-1979, this procedure resulted in annual lists of 10-14 eligible programs, which were then distributed to hospital volunteers ($N = 19-30$, mean age = 37-48 years), with the instruction "place a check mark beside the TV shows you have heard of." On the basis of responses to this check-mark procedure, the five most recognized programs were used to add five new multiple-choice questions to the test. All test items took the form "Which of the following was a TV show?" The correct answer was presented with three incorrect choices, randomly selected from a list of fabricated titles. In addition, items from the most remote time period were then calculated from the test, so that each updated version was limited to programs broadcast during the previous 15 years.

During the past 6 years, it became clear that the method originally employed for selecting one-season television programs (Squire & Slater, 1975) did not insure that programs selected for the test had in fact been broadcast only during the designated year. Thus, knowing that a program had premiered in October 1973 and was not scheduled during prime time in October 1974 did not necessarily identify it as a one-season program, since it might have in 1974 been scheduled outside prime time or it might

This work was supported by the Medical Research Service of the Veterans Administration, by NIMH Grant MH 24600, and by NIMH Clinical Research Center Grant 1P50 30914 01. We thank Paul Clopton for assistance with statistical analysis, Pamela Slater for research assistance, and Neal Cohen for helpful discussion. M. M. Fox is now at the Department of Psychology, University of Victoria, British Columbia.

have been broadcast later during the summer. It therefore proved necessary to consult the local newspaper or *TV Guide* throughout every year covered by the test to verify that each program appeared only in the designated year. With this procedure, 15 of the 75 programs originally selected for the test in 1973 were ineligible by 1976, either because they had been broadcast during the interval 1973-1976 or because they had been broadcast after the year in which they first appeared. In most cases, the additional broadcast time amounted to a summer following the year a program had originally appeared, and in no case was a program broadcast for more than 2 years. Then, using the checkmark procedure described above, these 15 programs were replaced with the next most recognized, eligible programs from the same year. However, this procedure required that eliminated programs be replaced on later tests by less well-known ones and would have complicated comparisons between tests. It was therefore decided that programs on later tests that became ineligible should be eliminated from the test but should not be replaced.

Nine additional programs have been determined ineligible and have been eliminated since the construction of Test 2 in 1975, eight from Test 1 and one from Test 2. Taken together, aside from the changes associated with adding and eliminating time periods, the updating procedures have resulted in 23 programs changed from Test 1 to Test 2, 1 program changed from Test 2 to Test 3, and no programs changed from Test 3 to Test 4. Thus, whatever bias might have been introduced into Test 2 to complicate its comparison with Test 1 could not affect comparisons among Tests 2, 3, and 4. Indeed, since Test 2 was constructed in 1975, it has seldom proved necessary to make any changes in test items. It is worth noting that these facts of test construction correspond to television programming practices. Before 1970, it was common for canceled programs to reappear in later years. Since 1970, perhaps because of greater competition for programming, canceled programs have reappeared infrequently.

These test construction procedures resulted in 67 questions for Test 1, given in December 1973, 67 questions for Test 2, given in November 1975, 68 questions for Test 3, given in November 1978, and 68 questions for Test 4, given in November 1979. Tests were administered on each of these four occasions to employees and volunteers at the San Diego Veterans Administration Hospital. For Test 1, the subjects (13 female, 4 male) averaged 54 years of age ($SD = 12$) and 12.6 years of education. For Test 2, the subjects (19 female, 13 male) averaged 40 years of age ($SD = 12$) and 13.8 years of education. For Test 3, the subjects (17 female and 15 males) averaged 40 years of age ($SD = 4$) and 14.4 years of education. For Test 4, the subjects (15 females and 11 males) averaged 47 years of age ($SD = 8$) and 14.0 years of education. All subjects indicated that they had had a television set in their households during the past 15 years.

RESULTS

Scores were grouped into five 3-year periods, and the results for all four tests are shown in Figure 1A. Because the four versions of the test were given on four different occasions from 1973 to 1979, the results from the four tests have been superimposed here, to permit comparison between events of similar age. Taken together, remote memory for the names of former one-season television programs was very good, and scores declined as a function of how long ago a program had been broadcast. Performance averaged 69.5% for programs that had been broadcast from 1 to 3 years prior to testing and 53.4% for programs that had been broadcast 13-15 years previously.

Figure 1. Recognition of the names of former one-season television programs as a function of the time since they were broadcast. Recognition memory was assessed on four different occasions: in 1973 for the period 1958-1972 (Test 1, $N = 17$), in 1975 for the period 1960-1974 (Test 2, $N = 32$), in 1978 for the period 1963-1977 (Test 3, $N = 32$), and in 1979 for the period 1964-1978 (Test 4, $N = 26$). For each test, means are shown for five 3-year time periods covering the 15 years sampled by the test. (A) Observed means. (B) Means after adjustment for the effect of the number of weeks programs were broadcast.

A two-way analysis of variance revealed a significant effect of time period on test scores [$F(4,412) = 25$, $p < .001$]. A trend analysis revealed a significant linear component of variance [$F(1,102) = 93$, $p < .001$]. Recognition performance declined linearly as a function of how long ago a program had been broadcast. Separate analyses for each test revealed a significant effect of time period in each case ($F_s > 5.6$, $p_s < .01$). Moreover, for each test the linear component of variance was significant ($F_s > 10.9$, $p_s < .01$).

The effect of test version was not statistically significant, indicating that the ability to recognize 1- to 15-year-old events was rather similar from test to test, 1973 to 1979. The detailed appearance of each curve differed to some extent, and the interaction of Time Period by Test Version was significant. Apparently, how long ago a program was broadcast was not the only factor influencing recognition, and programs from some time periods were recognized

better or worse than would have been expected solely on the basis of the time period to which they belonged. At the same time, it was clear that idiosyncrasies of particular programs were not a dominant influence on recognition, since irregularities in the curve from one test did not appear at appropriate points in the curves from other tests.

In order to be able to compare remote memory test scores from one time period with those for another, it must also be demonstrated that the test satisfies the criterion of equivalence, that is, that the items selected from different time periods were initially learned about to the same degree (Squire & Cohen, *in press*; Squire & Slater, 1975). Data relevant to the equivalence of test items come from extent of television viewing during the years covered by the test, individual Nielsen ratings that estimate the size of the national viewing audience for the programs on the test, and information about how many weeks each program was broadcast. Figure 2 shows percent of prime time viewing by individuals in television households since 1966. Time spent watching television has been remarkably stable. During the same period, the number of U.S. households with a television set increased only from 94% to 98%.

Individual Nielsen ratings, as determined by annual November rating periods, were obtained for most of the programs on each of the four tests (79% of programs on Test 1, 88% of programs on Test 2, 94% of programs on Test 3, and 95% of programs on Test 4). The Nielsen rating estimates the percent of U.S. television households watching a designated program. A two-way analysis of variance indicated that the programs selected from different time

periods for Tests 1-4 had reached similar-sized audiences [$F(4,221) = 1.9, p > .1$]. For all 20 time periods (4 tests $\times$ 5 time periods), the average Nielsen rating ranged from 13.0% to 17.8%. In addition, in part because of the overlap in programs from test to test, there was no difference among the four tests with respect to audience size.

Finally, by consulting microfilms of the local newspaper, we determined how many weeks each program on the tests had been broadcast in the San Diego area. Programs on Tests 1-4 had remained on the air an average of 32.8 weeks (range for means = 29-37 weeks). Separate one-way analyses of variance indicated that programs selected from different time periods for Tests 1 and 2 had been broadcast for similar lengths of time ($F_s < 2.0, p_s > .1$). Since about 1975, however, decisions to cancel programs have been made earlier in the season than in previous years, and for Tests 3 and 4, the time period to which a program belonged was significantly related to how many weeks it had been broadcast ($F_s > 4.8, p_s < .01$). This effect was due to programs in the most recent time period (1975-1977 for Test 3 and 1976-1978 for Test 4) having been broadcast for fewer weeks than programs in the more remote time periods. When the data for Tests 3 and 4 were analyzed excluding the most recent time period, there was no effect of time period on the number of weeks that programs had been broadcast.

The finding that the number of weeks a program had been broadcast was not equivalent across time periods for all the tests raised the possibility that this variable might have influenced recognition scores and thereby have affected to some degree the forgetting curves generated by the four tests. A multiple-regression analysis, using the three variables of Nielsen rating, weeks broadcast, and time period, revealed no significant effect of Nielsen rating on recognition scores, perhaps because of the restricted range of this variable, but yielded significant effects of the other two variables ($F_s > 20.0, p_s < .01$). Because Nielsen ratings were not available for a number of programs, this variable was then eliminated, and the regression analysis was repeated for all programs on the tests, using two independent variables, weeks broadcast and time period. The results were the same (for weeks broadcast, $F = 25.4, p < .01$, which accounted for 8.2% of the variance; for time period, $F = 30.8, p < .01$, which accounted for 10.0% of the variance). Next, the recognition scores for each time period and each test (shown in Figure 1A) were adjusted, using predicted mean scores for each test and each time period, as derived from the multiple-regression analysis. These adjustments were based on the overall mean for the variable weeks broadcast (mean = 32.8). Thus scores for time periods that consisted of relatively short-lived programs were raised (e.g., the most recent time periods for Tests 3 and 4), and scores for time periods consisting of relatively long-

Figure 2. Percentage prime time viewing during the period 1966-1978 for adult (18 years and over) male and female residents of television households in the continental U.S. During 1966-1972 prime time viewing data were drawn from the time period 7:30 p.m. to 11:00 p.m. each weekday. Since 1972, this period has been 8:00 p.m. to 11:00 p.m. The number of television households in the continental U.S. represented 94% of all households in 1966 and 98% of all households in 1979. Data were generously provided by the A. C. Nielsen Company.

lived programs were lowered. This procedure resulted in only minor changes in the forgetting curves generated by each test (Figure 1B; mean adjustment = 1.4%, range = .3%-4.3%).

DISCUSSION

Remote memory for past events was assessed on four different occasions during a 7-year period, using annually updated versions of a recognition test for former, one-season television programs. Each test yielded a forgetting curve that was primarily monotonic and linear, and each test's estimate of remote memory for the previous 15 years was rather similar to those of the other tests. The finding that the forgetting curve generated by the first version of the test, given in 1973, was largely recapitulated on three different occasions during the next 6 years provides strong evidence that this test can reliably measure remote memory.

As discussed previously (Squire & Cohen, in press), the major difficulty with available remote memory tests has been that it is difficult to determine whether they satisfy the criterion of equivalence, that is, that the items from different time periods were initially learned to a similar degree and that they were then forgotten at the same rate. The present findings suggest that the remote memory test described here satisfies these requirements.

Several factors that could have influenced degree of initial learning of the items in different time periods were examined directly. First, evening television viewing has been remarkably stable since 1966. Second, the Nielsen ratings for programs selected from different time periods were almost identical, and these ratings did not significantly affect recognition scores on the tests used here. Third, the number of weeks a program was broadcast was a significant factor in recognition scores, and this variable affected scores unequally across time periods, apparently because of the recent trend toward canceling unpopular programs earlier in the television season. However, adjusted recognition scores derived to correct for the influence of this variable caused only minor changes in the data, and the shapes of the forgetting curves were essentially unchanged by this procedure.

Taken together, the findings described here suggest that the television test can provide a useful and reliable measure of remote memory. Whereas a single application of the test can sometimes yield nonlinear effects, especially if small numbers of programs are used in each time period (Harvey & Crovitz, 1979; Levin et al., 1977), when a sufficient number of test items are selected by rigorous criteria, the test can yield reliable data year after year.

Whereas recognition memory for television program titles declined gradually during the 15 years covered by these tests, recognition memory for high school

classmates is reportedly preserved nearly intact for at least 15 years (Bahrck et al., 1975). The relatively slow forgetting observed in both of these cases sharply contrasts with the more rapid forgetting observed in typical massed-practice, laboratory learning sessions. Memory for television program titles, which is acquired by distributed practice during a period of several months, presumably constitutes an intermediate case between memory for material acquired in the laboratory situation and memory for high school classmates, which is acquired during a period of years.

In summary, repeated applications of a remote memory test during 7 years have recapitulated the forgetting curve obtained by a single retrospective application of this test. The findings strengthen the conclusion that this test satisfies the criterion of equivalence (Squire & Cohen, in press; Squire & Slater, 1975) and suggest that such tests can be a useful and valid method for a wide variety of clinical and experimental investigations of remote memory.

REFERENCES

- ALBERT, M. S., BUTTERS, N., & LEVIN, J. Temporal gradients in the retrograde amnesia of patients with alcoholic Korsakoff's disease. *Archives of Neurology*, 1979, **36**, 211-216.
- BAHRICK, H. P., BAHRICK, P. O., & WITTLINGER, R. P. Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, 1975, **104**, 54-75.
- BOTWINICK, J., & STORANDT, M. *Memory, related functions and age*. Springfield, Ill: Thomas, 1974.
- HARVEY, M. T., & CROVITZ, H. F. Television questionnaire techniques in assessing forgetting in long-term memory. *Cortex*, 1979, **15**, 609-618.
- LEVIN, H. S., GROSSMAN, R. G., & KELLY, P. J. Assessment of long-term memory in brain damaged patients. *Journal of Consulting and Clinical Psychology*, 1977, **45**, 684-688.
- MARSLER-WILSON, W. D., & TEUBER, H. L. Memory for remote events in anterograde amnesia: Recognition of public figures from newsphotographs. *Neuropsychologia*, 1975, **13**, 353-364.
- SANDERS, H. I., & WARRINGTON, E. K. Memory for remote events in amnesic patients. *Brain*, 1971, **94**, 661-668.
- SQUIRE, L. R., CHACE, P. M., & SLATER, P. C. Retrograde amnesia following electroconvulsive therapy. *Nature*, 1976, **260**, 775-777.
- SQUIRE, L. R., & COHEN, N. Memory and amnesia: Resistance to disruption develops for years after learning. *Behavioral and Neural Biology*, 1979, **25**, 115-125.
- SQUIRE, L. R., & COHEN, N. Remote memory, retrograde amnesia, and the neuropsychology of memory. In L. Cermak (Ed.), *Memory and amnesia*. Hillsdale, N.J: Erlbaum, in press.
- SQUIRE, L. R., & SLATER, P. C. Forgetting in very long-term memory as assessed by an improved questionnaire technique. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, **104**, 50-54.
- SQUIRE, L. R., SLATER, P. C., & CHACE, P. M. Retrograde amnesia: Temporal gradient in very long-term memory following electroconvulsive therapy. *Science*, 1975, **187**, 77-79.
- WARRINGTON, E. K., & SILBERSTEIN, M. S. A questionnaire technique for investigating very long-term memory. *Quarterly Journal of Experimental Psychology*, 1970, **22**, 508-512.

(Received for publication September 18, 1980;
revision accepted December 18, 1980.)