# How the ARCS[*] Was Done

## Introduction

NCSALL's Adult Reading Components Study (ARCS) was the first large-scale attempt to describe the reading of students enrolled in adult basic education (ABE) and English for speakers of other languages (ESOL) using a battery of individually administered reading and language tests. From May 1998 to June 1999 nearly 1,000 adult learners were tested at over 30 learning centers in eight states. This report recounts in practical terms how Adult Reading Components Study was carried out. Organized chronologically, it covers the initial design of the test battery and questionnaire, piloting, site selection, interviewer training, preparation of materials, interaction with participating adult education centers, scheduling, interviewing, and the procedures used for scoring incoming data and entering it in the data base. The intended audience for this report includes individuals and groups who are considering carrying out similar research, such as US DOE agencies, university-based researchers, and state and local adult education officials.

Because this was the first study of its kind, it was inevitable that a number of unanticipated obstacles would have to be overcome in the field. In assembling this report we have tried to present a thorough and balanced picture of our difficulties as well as our successes. Whenever possible we offer the lessons of hindsight so that future researchers will be spared some of our missteps.

## Expanding the ARCS

The ARCS was based on a previous study by Strucker (1995) in which the *Diagnostic Assessments of Reading* (DAR) (Roswell, F. and Chall, J.S., 1992), the *Test of Auditory Awareness Skills* (TAAS) (Rosner, J., 1975), and a brief questionnaire were given to 120 ABE students at five Massachusetts adult literacy centers. The 120 reading profiles were subjected to cluster analysis, which yielded nine clusters of adult readers ranging from beginners to GED-level.

The original intent of the ARCS as proposed to OERI in 1995 was to extend this methodology to a larger, more carefully sampled group of 400 ABE students. However, discussions in 1996 with OERI and OVAE focused on the desirability of an even larger study. As a result, expanded funding was provided by OVAE to carry out a study of 600 ABE students and an additional 400 ESOL students. The larger study would also include testing sites outside of the New England states.

## Design of the Test Battery

[*] The Adult Reading components Study was funded by the US Department of Education Office of Educational Research and Improvement (OERI) and the Office for Vocational and adult Education (OVAE). For more information please contact John Strucker via email at john_strucker@harvard.edu, or by telephone at (617) 495 - 4745

Based on the principles in the ARCS proposal, Strucker outlined the following criteria for selecting tests for the ARCS battery:

1. We wanted ABE and ESOL students to be able to complete our test battery and interview in one session, at their learning centers, and during the time they would normally be attending their classes. This meant that testing and interviewing had to be completed within two to three hours. We knew that if testing and interviewing ran longer than three hours, many students would require two sessions, and we would thus run into problems of missed appointments, schedule changes, and a few students even dropping out of school before the second part of their interviews could be completed.
2. Each test had to assess a skill that was known through previous research to be related directly or indirectly to reading comprehension, the ultimate purpose of reading. More specifically, we wanted to use *achievement testing* in reading; that is, tests of the components of reading (word analysis, word recognition, oral reading, and vocabulary) known to contribute to silent reading comprehension. (See Chall, J.S. and Curtis, M.E., 1991. "Diagnostic Achievement Testing in Reading." In Reynolds & Kamphaus, Eds., *Handbook of Psychological and Educational Assessment of Children.* NY: Guilford.) In short, we wanted to administer testing similar to what ABE and ESOL students would receive if they went to a reading specialist at a hospital or university reading clinic, or that they might have received when they were children from their K-12 reading teachers. In line with this approach, a small number of other assessments were included that test underlying processing abilities related to reading such as phonological awareness, short-term memory, and rapid automatized naming.
3. Our audience for the study included not only the research and policy communities, but also ABE and ESOL practitioners. Therefore, we wanted both the rationale for the testing and the tests themselves to be readily accessible to those practitioners.
4. Related to the preceding criterion, we also wanted our tests to be fairly straightforward to administer. Because we planned to use ABE and ESOL teachers as our interviewers as much as possible, the testing techniques had to be easy for them to learn with training of short duration. In addition, once the study was completed we hoped the results would encourage ABE and ESOL teachers to use some of these actual tests to help pinpoint the strengths and needs of their students.
5. If tests might eventually to be used by ABE and ESOL teachers on an every-day basis, they had to be relatively inexpensive for ABE programs to purchase.
6. We wanted the tests to be suitable for adults in terms of the content of the items tested.

7. We wanted to be able to score most of the reading tests during the actual testing session so that the students who were being tested could be given some immediate oral feedback on their strengths and needs.
8. Because our testing time was limited, we wanted tests that provided multiple sources of data. For example, the DAR Silent Reading test not only assesses comprehension with multiple choice questions, it also asks the student to give a brief oral summary of each passage he reads, thus providing samples of expressive oral language

With these criteria in mind, during 1996-97 Principal Investigator John Strucker, Assistant Director Ros Davidson, and ESOL consultant Ann Hilferty reviewed a number of reading and language tests and batteries. We also consulted many colleagues engaged in reading research, especially those with reading clinic experience, including: Marilyn Adams, Jeanne Chall, Carol Chomsky, Mary Beth Curtis, Rebecca Felton, Charlie Haynes, Pamela Hook, Vickie Jacobs, Steven Reder, Catherine Snow, Joseph Torgesen, and Marianne Wolf.

Almost immediately we were forced to reject the ABLE and TABE reading comprehension tests, even though they had been extensively normed on the ABE population. Both of these tests are widely used by many literacy centers and by state ABE administrators for monitoring student progress; therefore, we were concerned that many of the students we planned to test might have taken either test recently. The CASAS and TALS functional tests have also been extensively normed on adults; however, they are also becoming widely used in the field. Connecticut, for example, uses CASAS on a statewide basis to track student progress.

For the assessment of reading components, we were left with either the Woodcock-Johnson family of tests or the Diagnostic Assessments of Reading. The Woodcock-Johnson tests have been extensively normed. But they are more expensive, more time consuming to administer and score, and somewhat less “user-friendly” for interviewers who are not formally trained in assessment. The DAR was developed clinically and was therefore not as widely normed as the Woodcock, but it is easier to use and much less expensive than the various Woodcock-Johnson batteries. In addition, because the Woodcock batteries are more time consuming to score, tester would not be able to give feedback to students immediately following testing as described in criterion 7 on the previous page. Piloting both alternatives in 1996-97 (described below) ultimately led to the selection of the DAR as the primary English reading battery for the ARCS. However, the Woodcock-Johnson Word Attack subtest was used for English testing, and Spanish speaking students were assessed in Spanish reading using three parts of the Woodcock-Munoz battery.

Our greatest challenge was finding an assessment of English listening skills to use with ESOL students. The BEST and JOHNS are heavily used, but some ESOL practitioners have raised questions about their validity. With some reservations, we decided to use the listening comprehension section of the Language Assessment Battery

(LAB), a test designed by the New York City Board of Education to place ESOL and bilingual children in the appropriate types and levels of classes. The LAB assesses both conversational listening skills and more advanced listening skills associated with formal school-like situations.

Unfortunately, in the field the LAB was occasionally administered incorrectly by interviewers, more because of inadequate training than for any reasons intrinsic to the test. However, even when administered correctly, it appeared to separate ESOL learners into only three broad categories - those with little or no English listening skills, those with intermediate listening skills, and those whose listening skills were nearly as good as native speakers.

In retrospect, we believe that it was a mistake to use the LAB, whose norms are difficult to translate into meaningful adult categories. We might have achieved better results by simply reading aloud graded English passages of increasing difficulty to ESOL students and then asking them questions about them, similar to what Sticht and James recommend for the assessment of native English speakers (Sticht, T. & James, J.H., 1984 “Listening and reading” in Pearson, D., Editor, *Handbook of reading research,* 293-317).

## Drafting the questionnaire

We can certainly confirm the advice given to us by other researchers that the challenge of questionnaire writing is in limiting what to include. Especially in fields such as ABE and ESOL that have not been adequately researched, the temptation is to delve into many important areas such as employment history, motivation and persistence, health, parent-child relationships, and networks of support. However, because we knew that other NCSALL projects were focused more exclusively on some of these areas, we tried to limit our questionnaire to areas that were known through previous research to be directly concerned with reading. Even so, we found that we did not have time to ask every question that seemed important to us.

A team of four worked on drafting the questionnaire: the Principal Investigator Strucker, Assistant Director Davidson, ESL Consultant Ann Hilferty, and a qualitative research consultant, Christine Herot, who advised us on phrasing, organization, and interviewer directions. Our first step was to block out the major areas to be covered in the interview. We settled on the subject’s childhood home literacy environment, educational history, language history (for those who were not native speakers of English), history of reading disabilities (if any), self-assessment of reading strengths and needs, adult home and work literacy practices, reasons for pursuing adult education, and goals after completing adult education. In addition, NCSALL researcher Rima Rudd added several health and literacy questions. Lastly, before we went into the field, we received helpful feedback and several question suggestions from Darryl Mellard from NIFL’s National Center for Adult Learning Disabilities and the University of Kansas.

The four ARCS researchers worked on the wording of the questions together, going through four drafts before we ended up with the questionnaire that was ultimately piloted. Piloting (see below) resulted in further trimming of the questionnaire from 90 items down to 76 and rephrasing of questions that proved ambiguous or difficult for subjects to understand. After the final English version was ready, the questionnaire was translated into Spanish for use with beginning ESOL Spanish speakers.

**The creation of five testing protocols**

The mission of the ARCS was to assess and interview students from both ABE and ESOL classes. We assumed that we would be able to recruit interviewers who could interview and test in Spanish speaking ESOL students. But the range of other languages present in ESOL classes in the US proved to be quite daunting. According to the NALS, after Spanish, not one language out of the remaining 30 languages recorded in that survey accounted for more than 0.6% of the Level 1 and Level 2 population. It would have been impossible, or at least prohibitively expensive, to translate the interview questionnaire into all of the various languages we might encounter, and even more impossible to recruit and train interviewers who spoke these 30 or more languages. We briefly considered using advanced ESOL students as paid translators for some of the languages encountered, but rejected this idea because using students to interview students would have seriously compromised our guarantee of confidentiality to those interviewed.

Therefore, we were forced to limit our sample of ESOL learners as follows: We would test any and all Spanish-speaking ESOL students, from beginners who spoke little or no English all the way to advanced English speakers. But we would only be able to test non-Spanish speaking ESOL students if they spoke English well enough to be interviewed in English and to understand the test directions in English. In practice, this usually meant students in "intermediate" and above ESOL classes. To help us decide whom we could test, we always consulted teachers and administrators at the actual testing sites before deciding which of their ESOL classes to sample. We showed the testing materials to the teachers, described the testing and interview, and then asked for their judgment as to which classes of students they though we could test successfully.

After extensive analysis of the pilot results, we decided that we would need five different test protocols for the various categories of ABE or ESOL students we would encounter. All five of the protocols would include a core of English language reading tests, but additional tests would be given based on their appropriateness. (See Appendices 1-5, "ARCS Test Protocols.") All levels of ABE (including ASE and GED) and all levels of ESL would be tested in English reading skills using the DAR, Woodcock-Johnson Word Attack, and the Peabody Picture Vocabulary Test III (PPVT). Spanish speakers (whether in ESOL or ABE classes) would be tested in Spanish reading using parts of the Woodcock-Munoz Battery and in Spanish vocabulary using the Test de Vocabulario Imagenes Peabody (TVIP). Anyone whose primary language in childhood was not English was tested in English listening skills using the Language Assessment Battery. Tests of naming, phonological awareness, and short-term memory were translated and administered in Spanish to beginning ESOL students who were Spanish speakers. But ESOL students who were not native speakers of Spanish were not tested in these areas. Previous research and our own piloting of these materials with ESOL learners indicated that these tasks are very difficult to perform in a language that is not one's native language (or at least in a language not spoken fluently). Thus, difficulties with these tasks could not be taken as indications of underlying processing difficulties affecting reading.

**Assembling test packets**

The five test protocols were the basis for creating the five corresponding test packets. The protocols were also included in the individual test packets so that testers could use them as checklists to make sure they had administered all of tests needed for a particular learner.

Visiting scholar and ESOL consultant Ann Hilferty took on the critical responsibility for creating 1,000 individual test packets. This entailed making sure that each of the five different packets contained precisely the right materials assembled in exactly the right order. To accomplish this, Hilferty trained and supervised teams of graduate students to assemble the packets. Depending on the particular learner protocol, each packet contained between 30 and 40 individual pages of tests, questionnaire items, permission and payroll sheets, two DAR test response booklets, a blank cassette tape, and a sharpened pencil. All papers had to be three-hole punched, and each separate sheet had to be stamped with the subject's five-digit ID number. Colored stickers had to be attached to the appropriate pages to remind the tester when to turn the recorder on and off. All in all, the assembly of a single packet could take more than five minutes. But even then the work wasn't over. After assembly, each packet was rechecked to make sure that it was correctly done.

This painstaking work by Hilferty and her teams paid off enormously in the field. The labor invested in three-hole punching and applying stickers made the packets extremely user-friendly for the testers. Amazingly, in the entire ARCS covering a period of 12 months, there were fewer than ten instances when test packets arrived in the field missing needed items. Despite a complicated testing schedule in which we often had testers working simultaneously at three or more sites in three states, Hilferty was able to keep all testers supplied with the right materials, and we never had to hold up testing to wait for materials to be prepared.

We tried to learn in advance from teachers and administrators at the participating sites what kinds of learners we might expect at each site, but their estimates could not be precise, especially in distinguishing the various types of ESOL students. This meant that the packets could not be prepared very far in advance. Instead they had to be made in small batches on an ongoing basis depending on the types of learners at each site.

Keeping track of 1,000 packets was not easy. Some interviewers who traveled in their own cars and received their assignments over the phone were given assortments of packets to cover the range of learners they might encounter. Luckily, few packets were lost, and all but a handful were returned in a timely manner. Nevertheless, in future large studies like the ARCS, Hilferty recommends using a "tracking system" for packets. Each packet of test materials should be logged into the data base by its ID number *before* it goes into the field, recording the site to which it was sent, or the interviewer to whom it was given, and when it was returned, and whether it was returned completed or unused.

### Piloting Overview

The test batteries and questionnaire were piloted on 30 students from two adult literacy centers in the Boston area. Parts of the batteries and questionnaire were also piloted on an additional 11 students in the Harvard Adult Reading Lab. The piloting sample included ABE students who were native speakers of English (from beginners through GED levels), Spanish speakers enrolled in various levels of ESOL, and non-Spanish speaking ESOL students from intermediate and advanced ESOL classes. After each student in the pilot had been assessed, the researchers went over her or his testing and interview in detail, listening to tapes and re-reading notes.

### What was learned from the pilot study

As mentioned above, we confirmed that the RAN, Rosner, and WAIS Digit Span would only be useful when given in native language. This led to the translation of these tests into Spanish and our decision not to use those three tests at all with non-Spanish speaking ESOL students.

In the beginning of the pilot, the entire questionnaire was given at the beginning of the session before the start of testing. But it took twice as much time as we had allotted for it compared to when we had timed its administration ourselves. Most of the questions were designed for short answers. However, many students appeared to be so pleased to have a chance to talk one-on-one with a sympathetic interviewer about their education and their reading that they took 40 minutes or more to answer a questionnaire which was designed to take 20 minutes.

Two remedies suggested by Davidson were implemented:

- First, interviewers were trained in techniques for politely cutting off responses that were more detailed than we needed or responses that strayed off the topic.
- Second, we decided to give the questionnaire in two parts. Part A, lasting only 2-3 minutes, would provide the essential information the tester needed to administer the reading battery efficiently (subject's age, years of school completed, and native language). Part B, containing the remaining 50 questions on educational history and literacy practices, would be administered after all testing was completed.

Administering the bulk of the questionnaire after testing was completed proved very successful in cutting administration time. Students seemed less inclined to offer lengthy answers after they had worked on reading testing for 1 to 2 hours. Giving the bulk of the questionnaire after testing also meant that interviewers were unaware of students' self-reports of reading difficulties and other academic problems until *after* testing had been completed. As a result, reported presence or absence of reading difficulties could not influence how interviewers scored reading tests.

But there was a trade-off: the questions that asked the learner to describe her/his reading difficulties were now being asked *after* the assessments had been done. Although learners were given no feedback on the reading tests until after they were asked for self-assessments of their reading, it is possible that some students became more aware of their reading strengths and needs and made use of this knowledge in their self-assessments. Therefore, this possibility will have to be taken into account in interpreting answers to the self-assessment questions.

A major purpose of the pilot was to help us pare down our long list of assessments so that our entire interview would fit within the students' 2-3 hour time constraints. Reluctantly, we had to drop several very useful tests, which we mention here for possible inclusion in future studies.

- The Woodcock-Johnson information tests in Social Science, Natural Science, and Humanities provided excellent detail in areas that are directly related to the GED and other academic endeavors. But they took too long to administer and appeared to correlate well with the much briefer, but less specific WAIS-III R Information subtest.
- We also got interesting responses to the "Noun Test" devised by Catherine Snow and her colleagues. In one variation of this task, subjects are asked to define well-known words such as "knife" or "nose," and their definitions are scored as to relative strength in "decontextualized language." In another related task, subjects are asked to name as many different kinds of "knives" or "noses" as they can. Piloting these tests revealed some interesting differences in students' abilities to define known words, but because of time constraints, we decided that we would have to be satisfied with somewhat similar data from students' DAR Word Meaning definitions.

## Managing testing materials

We were concerned that interviewers would have great difficulty managing the dozens of sheets of paper that made up each test battery and questionnaire. This problem was addressed by having each subject's test packet pre-assembled (described earlier), with each sheet numbered with the student's ID number, all items packed in a manila folder in the order in which they would be used, and all tests and the questionnaire pages three-hole punched. Before testing began, interviewers were instructed to take all of the papers out of the manila envelope and place them in a three-ring binder in the order in which they had been packed and the order in which they were to be given. Thus, if the interviewers gave the assessments and questionnaire in that order, all items would be given in the proper sequence, no tests would be omitted, and the dozens of important sheets of paper would be less likely to be lost or misplaced. Interviewers were also trained to check off each test completed on the subject's protocol form. At the conclusion of testing, testers were trained to remove all materials from the binder and return them to the manila folder along with the tape recording of the session.

### Tape recording

Piloting helped us to select tape recorders that offered the best resolution of speech at the lowest cost (Sony Model TCM-59V). More importantly, piloting helped us to decide which parts of the testing and interview needed to be tape-recorded. During early piloting we recorded the entire sessions, but in the actual study, this would have been unnecessary and expensive for all 1,000 learners. We decided to limit tape recording to those parts of the test where we knew from experience testers might be most likely to make scoring mistakes, namely the Rosner TAAS, DAR Word Recognition, Oral Reading, and Word Meaning, and the Woodcock-Johnson Word Attack. We also wanted tape recordings of the parts of the session where the subjects' oral language itself constituted the data, such as the summaries of the DAR Silent Comprehension passages, their questionnaire responses, and their DAR Word Meaning definitions.

However, once we had decided not to record the entire session, testers had trouble remembering when to turn their tape recorders on and off. We devised a simple system to remind them: A test or questionnaire section marked with a green stick-on dot was to be recorded, and sections marked with a red stick-on dot were not to be recorded. To remind testers to time oral reading rate (which many seemed to forget), we used a single blue dot.

### Scoring criterion-referenced tests reliably: "When in doubt, keep testing"

The DAR Word Recognition, Oral Reading, Word Meaning, and Silent Reading Comprehension tests are criterion-referenced. That is, a student is given increasingly more difficult material until she or he fails to perform at mastery, usually defined as 70-75% of responses correct. To save time and to avoid discouraging the student, the interviewer usually stops testing each component as soon as she or he fails to achieve mastery. This means that the interviewer in the field must decide on the spot at what level mastery has been attained. With some tests this decision is easy: for DAR Spelling, words are either spelled right or wrong, and Silent Reading Comprehension is tested using a multiple-choice format. But, DAR Word Recognition, Oral Reading, and Word Meaning require the interviewer's real-time judgments. In Word Recognition and Oral Reading, allowances must be made for learners' regional and ESOL accents. In Word Meaning, many students give poor quality, borderline definitions even for words they know.* (See the *ARCS Training Manual,* pages 11-16, for examples of difficult to score responses in these three tests.)

If a tester's scoring criteria are too strict, she or he may stop testing a component before the student has a chance to achieve mastery, rendering the score for the component useless for the study. Piloting showed us that testers had to be trained to keep testing whenever they had the slightest doubt about whether a student failed to master a given level and/or when a student narrowly missed mastering a level. As an additional safeguard, all of these difficult-to-score tests were tape-recorded. This allowed specially trained scorers (who were subject to statistical reliability checks) at NCSALL to listen to the tapes to determine students' mastery levels according to uniform criteria. If our interviewers in the field tested enough levels, we could be reasonably sure that each student had been allowed to attain her or his highest levels of mastery on the DAR tests.

### Mistakes made during piloting

The bulk of the pilot testing and interviewing was done by Davidson and Strucker, who are both experienced reading clinicians and also the designers the batteries and questionnaire. They were assisted by two graduate students who were also specialists in adult literacy. Thus, all of the piloters were very familiar with testing and interviewing, and they understood how each procedure contributed to the overall research goals of the ARCS. However, once the ARCS was underway, at least half of the interviewers were ABE and ESOL teachers who had much less experience with reading assessment and almost no familiarity with research. If we had used a few ABE and ESOL teachers as pilot interviewers, we would have gained a better sense of which tests would present difficulties for them. We could have then provided more training or support for those tests, or, if necessary, eliminated them from the battery altogether. (This issue is discussed more in "Recruiting and training interviewers" below.)

### Recruiting and training interviewers

As mentioned above in our discussion of the test battery, it was the aim of the ARCS to train local ABE and ESOL teachers to do a substantial amount of the testing and interviewing. We felt that the study would have greater credibility among teachers if they knew that ordinary teachers had been intimately involved in it. Moreover, we wanted ABE and ESOL teachers around the country to feel that they could learn to use and interpret these kinds of reading tests themselves with relatively little training. In the end, about 40% of the tests and interviews were collected by teachers, including all of the tests from Texas, Tennessee, and New York and most of those from Connecticut.

* This in itself is important data that the ARCS will report on. When a student's definitions for known words are either vague or to narrow, reading comprehension is likely to be impaired. See Curtis, M.E. "Vocabulary Testing and Vocabulary Instruction" (1987) in McKeown and Curtis, Eds. *The Nature of Vocabulary Acquisition.* Hillsdale, NJ: Erlbaum Associates.

We decided not to attempt to recruit graduate students from local universities in those four states. Although graduate students would have brought a strong research perspective to the work, we were concerned that many of them might lack experience with the ABE/ESOL learners and might not interact well with them, the teachers, or the administrators. We did, however, employ some graduate students in New England where Strucker and Davidson were able to select people with the necessary experience and empathy.

We learned one important lesson concerning the use of teachers as researchers. Teachers who were selected by us with the help of local administrators were generally more successful at data collection than those selected by local administrators alone. It was not primarily a matter of difficulty in administering tests; most teachers we trained were able to learn this. It was more a matter of their intrinsic interest in reading. Good testing requires the interviewer to maintain genuine curiosity about a student's reading skills in order to remain focused on the details of the student's performance over the course of a two- to three-hour battery and interview. Those who tested simply to make extra money were generally less successful than those who regarded the ARCS as an opportunity to acquire testing skills and new insights about adult readers.

When we were able to interview prospective testers in advance we were able to judge whether they possessed an intrinsic interest in the subject, and we were able to politely discourage those who appeared to lack this interest. However, even in cases where we had had no influence over the selection of interviewers, we decided that we would accept all who were recommended to us, rather than risk ill-feelings on the part of local administrators and teachers by rejecting people they had recruited.

For most of the testing in Massachusetts, Rhode Island, and New Hampshire we used crews operating directly out of NCSALL in Cambridge. These crews included Language and Literacy graduate students from HUGSE (some of whom had ABE/ESOL experience), local ABE/ESOL teachers with demonstrated expertise in reading, and some interviewers who defied categorization: two novelists (both with adult education experience), a retired reading specialist, a middle school reading teacher, a landscaper (only available during the cold months on days when it didn't snow), a bookbinder, a secretary, a documentary film maker, and an unemployed Ph.D. in history. It was essential to have such a core group of interviewers who were *not* working teachers. Even part-time ABE/ESOL teachers were usually teaching precisely when they were most needed for ARCS testing - that is, during morning and evening adult classes.

We assumed that it would be relatively easy to find Spanish speaking interviewers for our NCSALL-based team, but we were wrong about this. Spanish speaking graduate students at HUGSE were much in demand and tended to be already fully employed on other research projects. Spanish speaking interviewers were plentiful in Texas, but not all who were trained had the intrinsic interest in reading mentioned above, and two of them had somewhat limited ability to administer our English reading assessments.

We knew at the outset that we could not rely completely on working teachers to collect all of the data in every location. For example, to guarantee students' confidentiality, teachers were not permitted to test students from their own programs. In large urban areas such as New York City, Houston, or Knoxville this presented no special difficulties because we were able to schedule teachers to test at sites where they did not work. However, this was not possible in smaller programs located in more isolated areas. In these cases we had to send in our teams of graduate students and other outsiders.

We underestimated the difficulty of retaining our NCSALL-based crew of researchers. Since we were only able to offer them part-time work, we experienced ongoing turnover as people left to take full-time positions. This meant that we were recruiting and training replacements throughout the data gathering period and not primarily at the beginning of the study, as we had hoped. In our remote locations (Texas, Tennessee, and New York) this problem was not as severe. In those areas the local teachers had up to two months to gather test data. Thus they were able to fit their ARCS testing and interviewing around their regular work schedules, collecting as few as one or two student interviews per week.

By contrast when we went to New England centers, we usually attempted to test all 40 or 50 students selected from each center within one to two weeks in order to cover as many sites as possible during 1998-99 and to minimize the disruption caused to each center. In these instances it was most efficient if we could send a full carload of five people, with each interviewer testing two to three students per day. Testers were pleased with this because they could earn up to $300 per day, and adult literacy centers were pleased because they were inconvenienced by our presence for only four to seven schooldays. However, because of bad weather, holidays and vacations, limited tester availability, and learners' complicated class schedules, we usually did not completely achieve this level of efficiency at most sites.

### Training sessions

We had hoped to complete training most of our testers just prior to going into the field and then to be done with training. In retrospect, we should have realized this was impossible. As mentioned above, because testers were part-time workers, even our NCSALL-based testers had to be constantly replaced and new testers trained. Future studies should assume that training will be an ongoing activity, as it was for the ARCS.

Strucker and Davidson[*] traveled to New York City, Tennessee, Texas, and Connecticut to train testers in those areas. New York City training took place in May '98, Tennessee in September '98, Texas in December '98, and Connecticut in April '99. We attempted to schedule training at each site approximately two weeks before testers went into the field so that they would not forget the details of the training. However, in

[*] All training was done personally by Davidson and Strucker to help ensure consistency in how the was data gathered.

Tennessee some testers were delayed in getting into the field by scheduling problems and a fire which damaged the Knox County Adult Education Center. To offset this delay, Strucker made a return trip to Tennessee to provide refresher training.

The details of the tester training are covered in the *ARCS Interviewer Manual*, which served as both the training curriculum and a reference manual for the testers to use after training was completed. The lengthy manual was supplemented by a one-page "Short List of Testing Procedures" that testers could refer to quickly during testing if they forgot how to administer a particular test. This "Short List" was invented by Carey Reid, one of the first interviewers we trained, for his own use during testing. Working from Reid's idea, Davidson devised an official "ARCS Short List" that conformed perfectly to the manual's more detailed directions.

It is our firm conclusion that tester training should have been more extensive than the 12-14 hours we provided. The tasks facing our trainees were Herculean. They had to learn to administer between 10-16 tests not only correctly, but in the ARCS's highly uniform way. This entailed being familiar with 10 to 16 different sets of detailed directions, time limits, and scoring techniques. While doing all this, they also had to manipulate a tape-recorder, turning it on and off at the appropriate times, and they had to be aware of whether students understood and were following directions. In addition, they had to administer a 76-item questionnaire, and when the interview was over, they had to provide sensitive and appropriate feedback to the students on their reading performance. All of this had to be done in the real world of adult literacy centers - in spaces that were sometimes uncomfortable, with interruptions from street noise and nearby classrooms, and under time pressure to finish the testing during the student's regular class time.

More training would have undoubtedly made the testers' work easier and their data more consistent. But the ARCS was not budgeted to cover this additional training cost. Future studies with comparable testing should probably allow at least 50% more money for training than we did, whether they are using teacher-researchers or graduate students. Our testers were paid $25/hour for training; so, to train 10 people at a site, it cost approximately $3,000 - $3500, excluding the costs of materials and travel and lodging for the trainers. In our first training sessions in New York City and Cambridge, testers were paid for their training before they began testing, on the condition that they complete 10 or more interviews. This was a mistake that became apparent immediately when several people dropped out before doing any testing at all, or after completing only one or two tests. We did not try to recover this training money, but following this testers were told in advance that they would only be paid for their training *after* they completed 10 or more interviews.

The last phase of training took place after interviewers went into the field. They were instructed to send their first two test packets in to Davidson immediately. She checked their first two test packets carefully from beginning to end and listened to each interview tape in its entirety. Then she called or emailed each new interviewer with corrections of any mistakes and suggestions on how to make her or his testing more efficient.

**Reflections on training**

Testers made more mistakes in testing than we anticipated primarily because we underestimated how difficult it would be for them to administer so many assessments. As we suspected at the outset, most mistakes occurred on the DAR. For DAR Word Recognition and Oral Reading, testers had to make decisions on the fly as to whether a student's pronunciation of a word was correct or not. As we expected, testers encountered difficulties with non-native speakers of English in trying to decide whether a pronunciation resulted from a person's accent in English (in which case it should be scored as correct) or whether a pronunciation resulted from an inability to apply rules of English decoding. As might be expected, some testers who were not native English speakers had even more trouble making decisions about whether English words had been pronounced correctly or not.

Testers also had difficulty deciding whether students' definitions on the DAR Word Meaning were correct. Sometimes this resulted from a student's inability to clarify ambiguous or highly contextualized definitions. Or, sometimes it resulted when testers failed to realize that a student's definition was actually acceptable, especially when the student's definition was somewhat unusual or unexpected. Testers who were not native speakers of English themselves experienced the most difficulty with these kinds of responses. One tester marked as wrong a student's definition of *ancient – "older people"* without asking the student to elaborate. Another scored an ESOL student's definition of *disturbance* as *"something that changes your quiet"* as incorrect.

In retrospect, we should have screened more carefully those interviewers who were not native speakers of English to make sure that they could score DAR Word Recognition, Oral Reading, and Word Meaning correctly. Perhaps those interviewers whose English skills were not sufficient for this very demanding English testing could have been teamed with English-only interviewers, with one person testing in only in Spanish and the other testing only in English.

Fortunately, most tester scoring errors did not result in lost data largely because we had repeatedly emphasized during training that they should test beyond mastery level if they weren't sure how to score students' responses. If they followed this procedure, Davidson and her team of graduate students back at NCSALL could tell by listening to the tapes what the proper mastery level should have been, and enter the correct score in the database. But in some instances testers stopped a DAR subtest too soon (usually because they had counted correct responses as errors) with the result that the student never had a chance to achieve his or her mastery. In these cases, we simply didn't have an accurate score for that test.

When this happened, we did not want to lose an entire subject because of only one incorrectly scored test out of a total of 12-16 assessments. We knew that the clustering methods we planned to use would be able to cluster subjects who were missing one test score with other subjects who had similar test profiles on the remaining 11-15 tests. Therefore, we decided that if a subject were missing only one DAR assessment, we would keep him or her in the database.[*] There was one exception: if the a subject was missing the DAR Reading Comprehension, we could not keep her or him because we planned to use reading comprehension as a dependent variable for many of our analyses.

Much of our training focused on the DAR assessments because they are inherently difficult to score (for the reasons discussed above and also discussed in pages 9 - 17 in the *ARCS Interviewer Training Manual*). Our focus on the DAR led us to overlook the amount of training needed for the Woodcock-Johnson and Woodcock-Munoz Word Attack Tests that both employed non-words in English and Spanish, respectively, to assess phonics knowledge. For people who are not experienced reading testers, it is not easy to keep track of whether a non-word such as *hopdalhup* or *pnir* has been pronounced correctly - especially if the subject reads too fast and/or speaks accented English.

Interviewers needed more time to practice this skill than we gave them in training, and they needed to tell students to read those words slowly enough for them to judge their accuracy. Fortunately, those tests were tape-recorded, so we have been able to listen to the tapes to establish correct scores.

**Processing and scoring finished tests**

Our "out of town" site coordinators in New York City, Texas, Tennessee, and Connecticut were asked to return completed test packets to NCSALL promptly, using prepaid FedEx labels billed to ARCS. In practice, this meant they returned completed packets in batches of 10-30 packets every few weeks. We urged them to be prompt because we couldn't pay students or testers until we had received completed packets.

Once packets were received at NCSALL, the Payroll/Permission Form was removed from each packet and logged into the database. Then student payment forms were forwarded to Harvard's Finance Office where checks were cut and mailed directly to students. Test packets were then checked to make sure that they were complete and that all pages contained the subject's ID number. If a test was missing or incomplete, we attempted to contact the interviewer immediately so that an additional session could be set up with that subject to gather any missing data.

The packets were then separated. Each test or scoring sheet was placed in a stack with the rest of its kind – e.g., all DAR booklets were placed together, as were all Woodcock-Johnson Word Attack scoring sheets, PPVT scoring sheets, etc. This was

[*] Naturally, in any presentation of clusters, we report how many subjects in that cluster had missing data.

done for efficiency's sake, so that the graduate students verifying scoring could work through stacks of the same tests. Tests that were easy for the interviewers to give correctly and easy to score required little attention by the graduate students. For the PPVT and WAIS III-R Digit Span, for example, the graduate students simply needed to check that they were correctly administered, count the responses to find the raw score, then turn to a table to look up the corresponding standard score based on the subject's age. The graduate student would then enter the subject's raw score and standard scores on a log sheet for that subject.

The log sheets listed all of the tests administered to the subject and some basic demographic information such as age, sex, program (ESOL or ABE), site where tested, years of childhood school completion, and native language. This information was listed in the same order in which it was laid out in the *Stata*© database to facilitate data entry.

The DAR tests and a few others needed to be carefully checked and verified. To ensure consistency, checking and scoring of these tests was done by only three people - Davidson and two assistants. All three scorers had extensive experience giving the tests themselves, and they met frequently before and during the scoring process to discuss scoring criteria with Strucker. Three separate inter-rater reliability checks were performed by having the three scorers each score the same twenty subjects' DAR tests. The first inter-rater reliability checks averaged .80-.90 (meaning 80 to 90% agreement among the three scorers). Later checks showed improved reliability, averaging above .95 across scorers.

As with testers in the field, Davidson and her assistants had the greatest difficulty deciding whether a non-native English speaking subject's pronunciations were reading errors, or the result of her/his accent. And, they also had difficulty deciding whether some subjects' vague definitions were correct or not.

We did not anticipate the problem Davidson called "interviewer drift." We had assumed that once a tester had received feedback on his testing from Davidson, and once his next few tests had begun to reflect that feedback, that all his subsequent tests would only need cursory checking and verification. To our surprise, this was not always true. After a few months of testing, a small number of interviewers appeared to "drift" away from the ARCS uniform procedures for administration and scoring and to develop their own slightly different approaches. As with other similar difficulties, "interviewer drift" was most likely to happen with people who had less intrinsic interest in reading and the purposes of the study. Fortunately, most of their mistakes could be corrected by listening to their tapes.

### Negotiating access to adult literacy centers

The first step in any field-based study is to negotiate permission from all of the people who will affected by the study. Especially in a decentralized system with considerable local autonomy such as ABE/ESOL, this usually involves more than a formal grant of permission from officials at the state level. When one negotiates permission to test students, one is also asking for cooperation that inevitably involves some inconvenience or at least a temporary change of routine on the part of the participants at all levels. Therefore, in asking people to cooperate in research one is always asking a favor.

In the case of adult literacy education, people in the field at all levels are over-worked, under-paid, and pulled psychologically in many conflicting directions. NCSALL researcher Carey Reid, a former ABE teacher and administrator himself, was only half-joking when he referred to adult literacy educators as suffering from "abused profession syndrome." With this in mind, ARCS negotiated entry as follows:

We asked Ron Pugsley of OVAE to mention our study to a meeting of state directors he attended in the fall of 1997. We then asked Pugsley to write a brief email to the state directors in the eight states we had selected for the ARCS, telling them to expect phone calls from Strucker asking if they would be willing to lend their support to the ARCS in their state. Preceding the phone calls Strucker had sent each a letter and a copy of the "ARCS Abstract." (See Appendices 6 and 7.) Strucker then called each state director, and, based on the time she or he could spend on the phone, explained the ARCS in detail. All eight state directors agreed to support the research.[*] Strucker then asked each to send a list of all of the ABE/ASE/ESOL Centers in their states.

The lists of programs supplied by state directors usually included what classes (ABE/ASE/ESOL) were offered by a given program and sometimes the number of students on roll in a year. We surveyed all programs in a state that served more than 100 students to inquire about the proportion of ABE versus ESOL and what native languages were spoken by the ESOL students. For reasons of economy, Strucker and Davidson decided not to include sites smaller than 100. About 10-15% of US adult literacy centers serve fewer than 100 students, but we had no reason to suspect that the students at smaller sites differed significantly with respect to their reading from those at larger sites.

This information was used by the ARCS sampling statistician, Tony Roman, to select from state lists programs that accurately represent the mix of students within that state. Strucker then sent a letter and a copy of the "ARCS Abstract" to the directors at selected sites explaining the study and requesting their participation.

Following this, Strucker made follow-up calls to each local site director. During these calls he explained the purpose of the study, how it was being funded, how the interviews would be conducted, what assistance the ARCS would require from them, and what other inconveniences the study might impose on their center if they chose to participate. He stressed to the local directors that their centers or their teachers were not

[*] The state director in Maine also readily agreed to lend his support to ARCS, but we were unable to test in Maine because we couldn't afford food and lodging for our NCSALL testers. In New York City we worked through the Literacy Support Initiative rather than at the state level.

in any way being evaluated and that the purpose of the study was only to describe the reading of students from around the US. He explained that our guarantees of confidentiality prevented us from disclosing to local teachers or anyone else information about an individual student.* However, if they desired a general report on the kinds of reading profiles present among their students or in their region, the ARCS could provide that after all data had been collected and analyzed. Strucker did not attempt to pressure the local directors; given the inconvenience imposed by ARCS, we did not want reluctant collaborators. Approximately 80% of all adult literacy programs contacted agreed to participate. The reasons programs gave for not participating included moving the center during the school year, lack of a director, or recent change in leadership or organizational structure, e.g., coming under the control of a new agency.

Given the difficulties in gathering statistics uniformly in the ABE/ESOL system, future researchers need to be aware that direct conversations with adult literacy center staff are the most accurate source of enrollment figures and program information. When Strucker spoke with the local directors, he learned that some sites had much smaller or larger enrollments than annual state figures had suggested. In some areas, any person who leaves his name at an introductory meeting is counted as an enrollee, even if he never attends classes. On the other hand, some programs had added whole new classes or programs since the last state figures had been compiled.

### Procedures followed at learning centers

Educational researchers in ABE/ASE/ESOL face a different set of conditions in the field from those encountered by K-12 researchers. K-12 researchers usually deal with established organizational hierarchies within school districts that are staffed by full-time professional administrators. In contrast, ABE/ASE/ESOL researchers frequently have to interact with over-taxed administrators, many of whom work part-time. While K-12 researchers can count on having 95% of the children present and available for testing on a given day, ABE/ASE/.ESOL researchers must contend with the much lower attendance and the highly transitory nature of the adult education population. The ARCS designed the following procedures with these and other realities of adult education in mind.

Once a learning center had agreed to participate in the study, a date was set to begin testing at that site. Before arriving to test, Strucker and Davidson arranged meetings at that center with the director and as many of the teaching staff as possible. These preliminary meetings proved to be vital to the success of our data collection. In

* Originally we had planned to allow students to sign an optional form giving permission for the tester to share what he or she learned about the student's reading with the student's teachers. But in the midst of data collection at our first site, we discontinued this practice. When we told a teacher (with the student's permission) that the student had difficulties with decoding and phonological processing, she said that we had confirmed her suspicions that the student was "heavily LD", she would therefore lower her expectations for his reading progress. Since we could have no control over what teachers would do with students' test information, we decided they would be better served if their individual results were not shared with their teachers.

the few instances when we were unable to hold preliminary meetings, we and the centers paid a price in confusion and minor misunderstandings that took precious on-site time to straighten out. There were several purposes for the meetings:

1. Directors and staff were briefed in detail on the testing and shown examples of the tests so that they would understand ARCS better.
2. Through frank discussions we learned which teachers were lukewarm toward the ARCS, perhaps because they distrusted testing. We then took extra time to try to explain to them how this testing would directly benefit teaching.
3. We took pains to assure teachers that were we were not investigating them or their centers - that we were interested in individual students' reading only. Davidson and Strucker further assured them that as experienced ABE/ESOL teachers we were certainly not going to think ill of them if some of the students we tested turned out to have severe reading difficulties.
4. Scheduling was laid out with vital input from teachers. If teachers were testing, taking class trips, or working in a different room from that on the printed schedule, we needed to know about it in advance so we could work around these minor glitches that could cost an interviewer 15 precious minutes on-site. Saving even small amounts of time was critical because our test battery was so long and we wanted to avoid having to schedule two sessions with a student.
5. Testing rooms were checked to make sure they were adequately lit, quiet, and contained desks and chairs.
6. Driving directions and driving times were worked out so that the research teams who followed could arrive in a timely manner for the testing. Teachers were also asked to recommend good local restaurants for visiting interviewers.

Once testing began, things generally ran smoothly – or as smoothly as can be expected under the normally difficult conditions of adult literacy. Because there are so many part-time teachers, communication within centers was not always perfect. An interviewer would arrive at a class (especially if it were remote from the main center) to find that the teacher had either never heard of the ARCS, or hadn't expected us that night. Interviewers and teachers generally handled these situations gracefully, but sometimes feathers were ruffled.[*] Not surprisingly, the testing tended to go more smoothly in centers that had experienced directors and adequate administrative staff. Smaller centers with over-worked teacher/directors tended to have more difficulty coping with the ARCS.

Any study of ABE/ASE/ESOL has to contend with what Tom Sticht has termed "student turbulence," meaning low and erratic attendance compounded by open-entry/open exit policies. Our original approach to sampling had called upon each center director to give us a list of all of the students in her school (excepting only those non-Spanish speaking ESOL beginners whom we could not test). From this list our sampling statistician would randomly select 30-50 names, depending on the center's size. We

[*] We suggest that future studies equip field teams with cell phones to facilitate solving such problems.

would then arrive at the school hoping to find those 30-50 students present on the days they were scheduled for class.

Disaster struck almost immediately. Five researchers would arrive at a site to find that of the 10 students expected for that time slot, only three were in school that evening. Some of the missing 7 were simply absent, some had changed nights, and others had dropped out in the two-week period between their selection and their testing. To our dismay, on lists of fifty selected students at a site, fewer than half of our subjects were actually getting tested. To make matters worse, we had to pay testers a $25 fee plus their travel expenses for trips when they found no students to test.

Very fortunately for the entire ARCS Jenelle Baker, our site coordinator in Houston, came up with a solution: rather than risking student no-shows by pre-selecting students, why not select participants from among students present on the day of testing by holding a lottery on the spot? We discussed Baker's suggestion at length with our sampling statistician, and implemented it in this form:

1. Before beginning testing at a site we asked directors give us a list of all classes, noting how many students had attended each class the previous week.
2. By knowing how many students were in each class and how many students were needed from the whole site, Strucker was able to specify how many raffle winners were needed per class. For example, from classes with 20 students, we might select two "winners," and select one "winner" from classes of 10 students. Thus, all classes would be sampled proportionate to their numbers, students would be selected blindly, and all students would stand the same chance of being selected.
3. Student participation in the ARCS, although paid, was completely voluntary and unpressured. Some students on winning the lottery would decline to participate. Interviewers conducting the raffles were asked to keep track of those declining and to record any reasons given for not wanting to participate to make sure that bias did not creep into the selection process. For example, were women declining in greater numbers than men? Were some ethnic groups more prone to decline than others?* As it turned out, most people who declined told us they did so because they had to leave early to pick up children or for some other practical consideration.
4. A pair of interviewers went into a class after the teacher had been warned several days in advance in advance of their visit. The teachers had been asked to announce the testing to their classes a day or two prior to testing by saying only that some researchers from Harvard would be coming to tell them about

* In one center, for example, females from Portugal appeared to be declining at a greater rate than other people. By talking to them, we learned that these women were somewhat reluctant to test alone with a male interviewer, and they were also under the misapprehension that ARCS testing would affect their school progress. Once aware of this, we had a female interviewer describe the study to the remaining ESOL classes, and in her description she made it clear that this testing was part of a national study that would have no effect on their school progress.

> some tests they were giving, that those chosen would be paid, and that participation was voluntary. We did not want teachers describing the study or attempting to "sell" it to the students. We wanted each person to decide for him- or herself whether or not to participate based on a reasonably uniform and neutral presentation of the ARCS. We didn't want teachers subtly suggesting that some people should participate instead of others, or expressing displeasure at the students if they did or did not participate.[ϕ]

As a general rule, in all of our dealings with teachers and administrators in learning centers we tried to be mindful of their feelings and perceptions of belonging to a field that is grossly under-funded and too often unappreciated by researchers and other outsiders. To help minimize the natural foreboding that some in the field might feel toward a Harvard/USDOE enterprise such as ARCS, Davidson and Strucker followed these guidelines in their interaction with administrators, teachers, and students:

1. We made it a practice to explain the ARCS in considerable detail to state directors, local administrators, program directors, teachers, support staff, and all students – not just those selected as subjects.
2. We wanted all those contributing to the research to feel that they could speak directly to the two people in charge of it. As a matter of principle, we tried to ensure that all initial and ongoing telephone conversations and face-to-face meetings with state directors, local administrators, program directors, and teachers took place directly with ARCS principal investigators Strucker and Davidson, even though our graduate students and staff could have handled some of these contacts. As the study progressed, student subjects were instructed to call Strucker directly if their payments were late, and teachers and center directors were instructed to call Strucker or Davidson directly with complaints or questions.
3. We tried to accommodate the preferences of literacy centers in terms of scheduling, even when that sometimes proved less convenient or efficient for our research teams.
4. In New York, Texas, Tennessee, and Connecticut we recruited respected local practitioners to coordinate research efforts in those areas. We solicited their advice on all matters pertaining to the data collection. Our site coordinators were paid stipends in addition to what they earned as ARCS interviewers.
5. The testing procedures were purposely designed to assess students at their *best* by attempting to provide conditions that would encourage them to perform at

[ϕ] In one class a teacher ignored the interviewer's emphatic request that she *not* describe the study to the students. She then immediately began to describe the study in somewhat negative terms, implying very strongly that she would be upset if a certain student missed her class that night because he already had a poor attendance record. When that student was among the raffle "winners," she glared at him, and he quite understandably declined. In a more positive vein, we often found ourselves having to turn down teachers' requests that we be sure to test a certain student "because she or he was very interesting or puzzling." We had to keep stressing that the selection of students had to be random in order develop a picture of the range of readers in the ABE/ASE/ESOL system, and that deliberately choosing "interesting or puzzling" students might distort this picture their own personal mastery levels. Oral reading fluency, for example, was measured on the highest level passage the student felt was smoothest. Silent reading time limits were generous. Testers were specifically instructed on how to make students feel comfortable and supported during testing without being patronizing or phony. Testers were also trained to give honest feedback at the end of testing, feedback that stressed students' strengths as well as their weaknesses in reading.

6. Students were paid $10/hour for their participation. We could not have asked them to miss an ABE or ESOL class for our interview without compensating them. This carried an unanticipated benefit: adult literacy directors and teachers told us that our paying their students led them to support the ARCS because it said to them that the ARCS respected their learners and took them seriously.
7. Students were guaranteed absolute confidentiality with regard to test scores and questionnaire information. They were also assured that although the US DOE was paying for the study, neither that agency nor any other government agency would have access to their names or social security numbers. They were also assured explicitly that their small payment ($20-$40) would not affect their income taxes nor would it lead to a reduction in any benefits they might be receiving such as welfare, disability, or SSI.
8. Participating students were informed at the start of the interview that they could discontinue the testing at any time for any reason and be still be paid for their time up to that point. In practice, only two people out of the entire sample of 1,000 chose to exercise that right, but we felt that students would feel more comfortable if they knew they were free to leave at any time.

**Unanticipated Costs**

It is not uncommon for any educational research to end up costing more than proposal designers anticipate. In the case of the ARCS, more accurate attention to costs during piloting might have helped. Specifically, if we had used less-experienced testers during the pilot, we would have known to allot more time and money for training. For the benefit of future researchers in ABE/ASE/ESOL, we have listed some of the areas where these additional costs occurred:

1. Tests and testing materials cost more than originally budgeted partly because we underestimated the number of testers (30 estimated, 50 actually) who would need to be in the field simultaneously, with each tester requiring two expensive items: a tape-recorder and a copy of the Peabody Picture Vocabulary Test.
2. We underestimated the cost of training testers in two areas. First, we underestimated the number of training hours they would require (6 estimated versus 10-12 actually needed). Second, we underestimated the total number of people we would need to train (at approximately $250 per tester) because

we didn't realize how much turnover would occur among our NCSALL-based testers.

3. We slightly underestimated how long the test batteries would take to administer, leading to slightly higher subject payments than we had anticipated.
4. At the outset, we assumed that testers would be eager to work for ARCS at a piece-rate of $50 per ABE/ASE student (based on two hours of testing time) and $75 for each Spanish-speaking student (based on three hours of testing time). But these rates proved to be too low. Given testers' travelling times and the fact that many tests took longer than the two or three hours per-test originally allotted, the above rate was simply not enough to attract and keep talented part-time workers. Therefore, when our New York testers complained that they were incurring high transportation costs and spending up to two hours of subway travel time, we raised the per-test payment to cover these factors. Out of fairness to testers at other sites, we made this our standard rate for all locations. We also decided to pay for on-the-road meals for NCSALL-based testers and an additional mileage reimbursement if they drove their own cars.

## Summary of key recommendations

For the benefit of researchers who are planning any studies of ABE/ASE/ESOL enrollees, we would like to re-state five recommendations that could help to make or break future studies:

1. Pay the ABE/ASE/ESOL student subjects (for reasons discussed above).
2. Test the students during their scheduled classes, using some form of on-the spot lottery.
3. Do not allow teachers or administrators to select or influence the selection of the students or classes that are sampled.
4. Always meet personally with teachers and administrators to explain the study in detail before beginning data collection.
5. Principal investigators should make themselves personally available to teachers, administrators, and students