Adult Reading Components Study (ARCS) Research Methodology

John Strucker, EdD
ARCS Principal Investigator

The Adult Reading Components Study (ARCS) (Strucker, J. & Davidson, R, 2003) is the source of the reading data that were used to construct the profile matching feature of Assessment Strategies and Reading Profiles (ASRP) website. Conducted under the auspices of the National Center for the Study of Adult Literacy (NCSALL),[1] the ARCS was the first large-scale study to describe the reading strengths and needs of students enrolled in classes for adult basic education (ABE), adult secondary education (ASE) 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 interviewed and tested at 24 learning centers in seven states.

The ARCS was based on a preliminary 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 instructionally relevant clusters of adult readers, ranging from beginning readers to those at GED level.

Although a total of 955 adult learners participated in the ARCS - 676 enrolled in ABE/ASE classes and 279 enrolled in ESOL classes - this paper focuses primarily on the ABE/ASE enrollees, because it was their data alone that were used to create the ASRP match a profile feature. For those who desire it, additional information on the data collection and analysis of the ESOL enrollees can be found in the Appendix I and in Strucker and Davidson (2003).

Test battery development

Each learner participating in the ARCS received an orally administered background questionnaire and a battery of reading tests. Several criteria were used in selecting the tests for the battery.

First, 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. We focused on *achievement testing* in reading; that is, tests of the components of reading (word analysis, word recognition, oral reading, and vocabulary) known to contribute directly to reading comprehension (Chall, J.S. & Curtis, M.E., 1990). The aim was 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 testing similar what children receive from K-12 reading specialists. In line with this approach, a small number of other assessments were included to test underlying processing abilities related to reading such as phonological awareness, short-term memory, and rapid automatized naming (RAN).

[1] The ARCS was funded by two offices of the US Department of Education: the Office of Educational Research and Improvement (now the Institute of Education Sciences) and the Office of Adult and Vocational Education, Division of Adult Education and Learning.

The intended audience for the study included not only the research and policy communities but ABE/ASE and ESOL practitioners as well. Therefore, the second test selection criterion was that the rationale for the testing had to make sense to practitioners. Moreover, because the study planned to use ABE/ASE and ESOL teachers as interviewers, the testing techniques had to be accessible to teachers who were not formally trained reading specialists. In addition, once the study was completed it was hoped that ABE/ASE and ESOL teachers around the country would come to view the ARCS testing approach as one they could readily learn and adapt for use with their own students.

The third test selection criterion was that the ARCS field researchers needed to be able to finish the interview and test battery in one session, and during the times students would normally be present to attend their classes. This meant that all testing and interviewing had to be completed within two to three hours. The final criterion was that the tests needed to be suitable for adults in terms of the content of the test items.

With these criteria in mind, during 1996-97 the investigators reviewed a number of reading and language tests and batteries. We also consulted colleagues engaged in reading research, especially those with reading clinic experience including: Marilyn Adams, Jeanne Chall, Carol Chomsky, Mary Beth Curtis, Rebecca Felton, Charles Haynes, Pamela Hook, Vickie Jacobs, Steven Reder, Catherine Snow, Joseph Torgesen, and Marianne Wolf.

We rejected the ABLE (1986) and TABE (1987) reading comprehension tests, even though they had been extensively normed on the ABE/ASE population because those tests were widely used by literacy centers and by state ABE administrators to monitor student progress, making it likely that many of the students we planned to test might have taken either test recently. The CASAS has also been extensively normed on adults, but it is widely used in the US, including in Connecticut, one of the seven ARCS states.

We were left with either the Woodcock-Johnson (1987) family of tests or the Diagnostic Assessments of Reading (1990), which both include assessments of the components of reading and reading comprehension. The Woodcock-Johnson tests have been extensively normed, but they are more expensive, more time consuming to administer and score, and less "user-friendly" for interviewers who are not formally trained in assessment. The DAR was developed clinically and was not as widely normed as the Woodcock, but it is easier to use and significantly less expensive than the latter. Piloting both alternatives in 1996-97 (described below) ultimately led to the selection of the DAR as the primary English language reading battery for the ARCS. The DAR was used for Word Recognition, Oral Reading, Spelling, Word Meaning (expressive vocabulary), and Silent reading Comprehension. However, the Woodcock-Johnson Word Attack subtest was used for English testing, and all Spanish speaking students were assessed in Spanish reading components using three subtests of the Woodcock-Muñoz (1981) battery.

The greatest challenge was to find a quick assessment of English listening skills for use with ESOL enrollees and non-native speakers of English enrolled in ABE classes. We decided to use the listening comprehension section of the Language Assessment Battery (LAB) (1982), 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. We had some reservations about the LAB because it assesses only listening and because some of the basic level items deal with childhood situations such as birthday parties.

Testing protocols

The mission of the ARCS was to assess and interview the range of students who were enrolled in ABE/ASE and ESOL classes – an extremely diverse learner population in terms of their native language backgrounds. To capture some of this diversity of reading and language skills, the ARCS employed five separate testing protocols. (See the Appendix II for all five protocols, including those used for the ESOL part of the ARCS.) For ABE/ASE enrollees, the following three protocols were used:

> Protocol 1 - for native speakers of English, consisting of English reading assessments only;
>
> Protocol 2 - for native Spanish speakers, consisting of both English and Spanish reading assessments and the Language Assessment Battery (LAB);
>
> Protocol 3 - for all other non-native speakers of English, consisting of English reading assessments and the Language Assessment Battery (LAB).

ARCS Protocols Administered to ABE/ASE Enrollees

Test Names (See References for complete test publishing information.)	(1) Native English Speakers	(2) Native Spanish Speakers	(3) All Other Nonnative Speakers of English
Diagnostic Assessments of Reading: Word Recognition, Word Analysis (consonant sounds), Oral Reading (level and rate), Word Meaning, Spelling, and Silent Reading Comprehension	√	√	√
Peabody Picture Vocabulary Test III	√	√	√
Wechsler Adult Intelligence Scale-III: Information and Digit Span	√	√	√
Woodcock Reading Mastery: Word Attack	√	√	√
Rosner Test of Auditory Analysis Skills	√	√	√
Rapid Automatized Naming	√	√	√
Test de Vocabulario en Imagenes Peabody		√	
Woodcock-Muñoz Bateria: Identificación de letras y palabras; Análisis de palabras; Comprensión de textos		√	
Language Assessment Battery		√	√

It was decided to test the native speakers of Spanish enrolled in ABE/ASE classes in both Spanish and English literacy skills in order to explore the range of native language literacy skills in the ABE/ASE population and the relationship of native language to second language literacy. We also wanted to identify similarities and differences between ABE/ASE Spanish speakers in both languages and the Spanish speakers enrolled ESOL classes. The availability of the Woodcock-Muñoz battery and Spanish speaking interviewers made this possible. Unfortunately, comparable tests and interviewers were not available to cover the remaining100 native languages represented in ABE/ASE and ESOL population.

Questionnaire construction

A team of four investigators worked on drafting the background questionnaire: the Principal Investigator John Strucker, Assistant Director Rosalind Davidson, ESL Consultant Ann Hilferty, and qualitative research consultant, Christine Herot, who

advised us on phrasing, organization, interviewer directions, and coding. The major areas to be covered in the learner interview included:

* childhood home literacy environment (including parents' educational completion);
* k-12 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;
* home and work literacy practices;
* reasons for enrolling in adult education;
* and goals after completing adult education.

In addition, NCSALL researcher Rima Rudd (personal communication, 9-8-97) added several health and literacy questions, and we received helpful feedback and several questions from Darryl Mellard at the University of Kansas (personal communication, 9-15-97). Piloting (see below) resulted in trimming the questionnaire from 90 items down to 76 and rephrasing of questions that proved ambiguous or difficult for participants to understand. The questionnaire was translated into Spanish to allow it to be administered to beginning ESOL Spanish speakers.

Pilot Study

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 administered to an additional 11 students in the Harvard Adult Reading Lab. The pilot 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 interviews in detail, listening to tapes and re-reading note**s.**

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.

A major goal of the pilot was to pare down the long list of assessments so that our entire interview would fit within the 2-3 hour time constraint. Reluctantly, we had to drop several very useful tests, two of 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 instructionally specific WAIS-III R Information subtest.

• We also got interesting responses to the "Noun Definition 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 "bicycle," and their definitions are scored as to relative strength in "decontextualized language."

Piloting helped us to select tape recorders that offered the best resolution of speech at the lowest cost (Sony Model TCM-59V) and 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 955 learners. We decided to limit taping to those parts of the battery 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 recorded those of the parts of the session where the subjects' oral language itself constituted the data, such as their questionnaire responses, and their DAR Word Meaning definitions.

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. 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.

For the majority of the testing in Massachusetts, Rhode Island, Connecticut, and New Hampshire we used crews operating directly out of NCSALL in Cambridge. These crews included Language and Literacy graduate students from Harvard (some of whom had ABE/ESOL experience), local ABE/ESOL teachers with demonstrated expertise in reading assessment, and some interviewers who defied categorization: novelists with adult basic education experience, a retired reading specialist, a middle school reading teacher, an ABD in linguistics, 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 because ABE/ESOL teachers were usually teaching precisely when they were most needed for ARCS testing.

Interviewers received 12-14 hours of training on the administration of the ARCS battery and Background Questionnaire. We had hoped to complete training most of our testers just prior to going into the field and then to be done with training, but this was not feasible. As mentioned above, because testers were part-time workers, testers had to be constantly replaced and new testers trained. For consistency, all interviewers were trained by Strucker and Davidson. The details of the training are covered in the ARCS Interviewer Manual (Strucker, Davidson, & Reddy, 2001), which served as both the training curriculum and a reference manual for the testers after training was completed. The lengthy manual was also 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.

Once the study began, Assistant Director Rosalind Davidson monitored the interviewers' work closely for fidelity and reliability. She reviewed their first two tests in detail and listened to their tapes, then gave each interviewer feedback on their accuracy as well as any tips for improving the efficiency of the testing process. Davidson also made period checks of all interviewers' tests throughout the study to maintain data quality.

Site selection

To create a sample of learners from urban and rural backgrounds that included African Americans, Hispanics, and whites, as well as a cross-section of English language learners, the following seven states were selected from which to recruit sites: Massachusetts, Connecticut, Rhode Island, New Hampshire, New York, Tennessee, and Texas. With the cooperation of US ED Office of Vocational and Adult Education (OVAE), researchers contacted the ABE directors of each state, all of whom readily agreed to assist in the study. The state directors furnished lists of all the programs within their states, including the previous year's enrollment totals, ethnic composition, and class schedules. The ARCS sampling statistician then drew up sampling frames for each state, specifying which programs to recruit and in order to obtain a representative cross-section of that state's urban/rural and ethnic mix. This procedure also allowed the researchers to provide any states who wanted them with informal reports about the learner profiles in their states. Local program directors were telephoned by Strucker, who explained the study to them and asked if they would like to participate. Over 80% of the programs contacted in this way agreed to participate.

Learner sampling

Initially the ARCS sampling statistician attempted to randomly select learners to be interviewed and tested from enrollment lists provided by each site. Almost immediately, this procedure proved to be unworkable because enrollments were very unstable. In the week between their random selection from the list and actual testing, significant numbers of students who had been selected had dropped out, and many who remained were often absent on the nights or mornings they had appointments to be tested. Interviewers frequently drove several hours to a site only to find that less than half of the students they were scheduled to test were actually present.

Janell Baker, the ARCS site coordinator in Houston, proposed that we select the learners for testing by a lottery conducted on the spot just minutes prior to testing. Taking Baker's suggestion, the researchers worked with the sampling statistician to devise procedures for conducting real-time classroom lotteries that met our criteria for random selection. These techniques were used with great success for the remainder of the study. All classes at a site were sampled proportionate to their size, with one participant selected for every 10 enrollees. The result was a reasonably random sample of learners that was not influenced by teachers' or administrators' decisions or opinions.

Most students selected via the lottery readily agreed to participate and were paid $10 per hour for their time. A record was kept of those who were selected but refused to participate to ensure that there were no patterns of refusal based on age, gender, or ethnic background, etc., that might affect the study's results.

Scoring the assessments

Test site coordinators returned completed test packets to NCSALL at Harvard promptly, using prepaid FedEx labels billed to ARCS. 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 participant to gather any missing data.

Rosalind Davidson trained and supervised teams of Harvard graduate students to score the various tests in the battery. The tests that were relatively easy for the interviewers to administer correctly in the field (the PPVT, WAIS Digit Span, or RAN) were also relatively easy to check and score back at NCSALL. However, the DAR tests

and a few others needed to be closely checked and verified. To ensure consistency, checking and scoring of these tests was done by only three people - Davidson and two graduate student 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 on these three scorers over the course of the study. Early checks averaged .80-.90 reliability, with later ones showing near .95. As with the testers in the field, Davidson and her scoring assistants had the greatest difficulty deciding whether a non-native English speaking participant's pronunciations were reading errors, or the result of her/his accent. And, they also had difficulty deciding whether some participants' often vague DAR Word Meaning definitions were correct or not.

Most of the field interviewers' testing mistakes on the DAR assessments could be corrected by the scorers by listening to the tape recordings. This is because the interviewers had been trained to keep moving forward to higher levels on these tests if there was any doubt whether a learner had reached mastery. Therefore, even if an interviewer in the field had under-estimated a learner's mastery on DAR Word Meaning at say grade equivalent (GE) 6, because all interviewers were trained to test a few additional higher levels, if that learner was actually found to have mastered GE 7 or higher, that determination could be made by the scorers listening to the tape.

Database construction

Participants' test scores and questionnaire responses were entered into either the ABE/ASE data base or the ESOL data base, depending on which program they had been enrolled in on the day they were tested. A total of 676 students were enrolled in ABE/ASE and 279 in ESOL. Later 202 of the 279 ESOL learners who were native Spanish speakers, and for whom we had collected Spanish literacy data, were placed in a separate data base for additional analyses.

Inevitably in a large study with multiple assessments there were some problems with missing or unusable scores, either because interviewers forgot to administer an assessment or because a testing error rendered the results of that assessment invalid. Of the 676 ABE cases, 457 were found to be totally complete after all the scoring had been verified.

It was decided provisionally to include any cases that were missing only one of the reading tests, as long as the missing test was not DAR Silent Reading Comprehension, the ARCS's only measure of comprehension. Of the 219 incomplete ABE cases, 21 were excluded because they were missing more than one reading test and/or DAR Silent Comprehension, leaving a total of 655 cases that were potentially suitable for inclusion in the study.

Following this, an analysis of the effects of statistically imputing the missing test data in the 655-member data set was conducted. After transforming the remaining 655-case raw test scores into standard scores, the missing data were imputed via Systat©. We then ran exploratory analyses and compared results between the imputed-plus-complete (655) and complete-only (457) cases.

1. First, we determined that the 219 cases with missing data were similar to the 457 complete cases in terms of their distributions for gender, age, ethnicity.

2. We next compared the correlation matrices for the imputed 655-case and complete 457-case data bases and found no significant differences.

3. Stepwise regression against DAR Silent Comprehension of the various components tests also revealed no differences in results between the imputed set and complete data sets.

4. No differences emerged between the imputed and complete sets in the overall configuration of the exploratory cluster solutions in shape or elevation, using both Wards and K-Means clustering methods (see below).

5. There was less than a 5% difference in actual cluster case membership between the imputed and complete data sets as a result of those individuals present in both sets having "migrated" to different clusters. [This was similar to the results of a later "hold-out" clustering procedure in which a random sample of 250 cases was held out of an initial cluster analysis, and then added back and subjected to a second cluster analysis.]

Based on the above comparisons, it was decided to use the statistically imputed larger 655 ABE case data set in the analysis in order to get the benefit of the most inclusive sample possible.

Cluster analyses

Cluster analysis is a statistical technique for grouping data - in the case of ARCS individual reading profile scores - according to similarity. Cluster analysis was chosen as the primary means of analysis for the ARCS for several reasons.

When investigators are looking at variables that do not have a substantial body of research and theory to connect them to each other or and a dependant variable, cluster analysis is at best exploratory – useful for generating hypotheses for testing. However, as in many public health studies, where the relationships among the variables are supported by theory and research, cluster analysis can help to identify useful profiles, subtypes, or syndromes. In the case of heart disease, for example, subgroups within a patient population might be identified by the degree to which they possessed certain known risk factors such as hypertension, smoking, high cholesterol, obesity, etc. In the case of the components of reading and their relationship to reading comprehension, although it may not rise to medical standards, strong theoretical and research connections have been established about their relationships to each other and to reading comprehension. (See for example, Carver, 2001; Gough & Hillinger, 1980; and Perfetti, 1985.)

Reading clinicians have long used reading profiles made up of the components of reading (Chall, J.S. and Curtis, M.E., 1991; Roswell, F. & Chall, J.S. 1994) to inform instructional decisions. Similarly, the aim of the ARCS was to describe the various reading profiles of ABE/ASE and ESOL enrollees in order to inform what placement and instruction in adult literacy centers. Prior to conducting the cluster analyses of the ABE data set, several exploratory steps were taken. First, the correlation matrix for all of the tests was examined to see which test variables correlated at low levels with DAR Silent Reading Comprehension because reading comprehension is the primary purpose for reading. Second, step-wise regression analyses were conducted, regressing all other tests against DAR Silent Reading Comprehension to see which tests appeared to contribute least to variance in comprehension. Further, although five versions of Rapid Automatized Naming were administered (letters, numbers, colors, common objects, and mixed letters/colors), it was found that all correlated highly with each other, and that

RAN letters correlated highest with DAR Silent Reading Comprehension. Accordingly, for the sake of reporting simplicity, it was decided to use only RAN letters in the analysis.

As a result of these investigations, it was decided to use the following tests in subsequent cluster analyses: DAR Silent Reading, DAR Word Meaning, DAR Oral Reading (both Grade Equivalent mastery level and rate in syllables-per-minute), DAR Word Recognition, DAR Spelling, WAIS Information subtest, PPVT, RAN Letters, Woodcock-Johnson Word Attack, and the Rosner Test of Auditory Awareness Skills (TAAS). Omitted from cluster analysis were the remaining additional RAN tests and WAIS Digit Span.[2]

Following earlier work by Strucker (1995), two clustering algorithms were used on the ARCS ABE/ASE data base, Ward's and K-Means. Euclidean Distance was used as the distance metric for both because it does the best job of capturing both the shape and elevation of clusters (Lorr, 1979). Using two very different clustering algorithms, such as Ward's and K-Means, is a way to verify that the cluster results are not simply artifacts of a particular algorithm; that is, if the two different algorithms produce highly similar results, then it is unlikely that the choice of algorithm is driving the results.

Ward's method was employed first. It is an agglomerative clustering algorithm that is reported as a dendrogram (stem and root) array. Each person begins as a "cluster of one" at the left- hand side of the diagram, and then each is merged with others in successive stages until at the right side of the diagram all are joined in one all-inclusive cluster. At each stage individuals are joined to others based on similarity. This allows Ward's to provide a visual representation of how similar various individuals or clusters are to each other and to note at what points outliers are joined to other individuals or clusters. Outliers can easily be identified as those who are the last to join clusters at any level. With Ward's method the researcher must decide what level or number of clusters is the most useful; in the case of ARCS this meant some point between the 676 single-member "clusters" and the final unitary 676-member cluster.

With K-means, the number of clusters to be created is specified in advance by the researcher. Accordingly, the K-means assigns every member of the data set to one of the specified number of clusters, and there are no outliers.

Cluster specification based on the criterion of "instructional relevance"

In the ARCS, the central criterion for determining the number of clusters to report was instructional relevance. In an ideal world where all learners could be tested with the entire ARCS battery and where highly trained teachers might have the luxury of working with individuals or small groups, 20-25 ARCS clusters might be considered "instructionally relevant." For example, at the second level of the Ward's cluster dendrogram, learners at the low intermediate ABE level whose native language was primarily Haitian Creole were placed in a separate cluster from another group of learners at the low intermediate ABE level whose native language was Spanish. At the next level of the dendrogram, these two clusters were joined and augmented by outliers from a

---

[2] Based on clinical experience and the sample distribution of Digit Span, we expected to find severe Digit Span difficulties concentrated among the most severely reading-impaired learners in the sample. Subsequent cluster analysis confirmed this: The two clusters comprised of reading disabled beginners had mean working memory scores 2-3 SD's lower than the learners in the remaining eight clusters.

variety of other native languages. In deciding to focus on this next level of the dendrogram (in effect lumping the Haitian Creole and Spanish clusters together), we reasoned as follows: In the real world of ABE, it is unlikely that most programs would be able to create two separate low intermediate ABE classes, one exclusively for their Haitian Creole speakers and another exclusively for their Spanish speakers. Moreover, even the best-trained ABE teachers would probably not significantly change their instructional approach for these two groups of learners in terms of the mix and levels of component skills in vocabulary, oral reading, spelling, and comprehension.

After examining Ward's clustering results in detail at various levels and applying the above criterion of "instructional relevance," the researchers decided to slice into the dendrogram at the third level from the left, which resulted in 10 clearly defined clusters of ABE readers.

The next step was to employ K-Means clustering, a method by which the number of clusters desired is specified in advance and in which there can be no outliers: all cases must be assigned to a cluster. Based on the above Ward's analysis, K-Means trial cluster solutions of 8, 9, 10, 11, and 12 clusters were created. Examining the means and distributions for the 11 reading assessments in the 8 through 12-cluster solutions and applying the criterion of "instructional relevance," it was decided that the 10-cluster K-Means solution would also provide the most useful and instructionally relevant description of the range of ABE readers, from GE 0-2 beginners through advanced GE 12 GED learners.

The next step was to show the 8, 9, 10, 11, and 12 K-Means cluster solutions to two experienced reading clinicians and two ABE reading teachers who were not connected to the ARCS to ask for their opinions on the most instructionally useful number of clusters. Although the two ABE teachers that 8 clusters might be the maximum practical number for the ABE system to handle based on its limitations, none of those consulted argued that we should report more than 10 clusters.

Reliability checks

Since Ward's and K-Means are entirely different clustering methods, a comparison between the two can be used to establish whether the clusters that have been created are simply driven by the clustering methods themselves. First, it was determined that mean standardized test scores from matching Ward's and K-Means clusters were sufficiently similar using t-tests, suggesting that shape and elevation were similar for both sets of clusters. Second, cluster memberships were compared across matching Ward's and K-Means clusters. Most matching clusters had 90% or better membership overlap, and none had less than 80% identical membership. Considering that all outliers were forced into the 10 K-Means clusters, whereas the most severe outliers had not yet been placed in Ward's clusters at the third or 10-cluster level, this 80-90% overlap was viewed as further evidence that the solution was not primarily the result of one or the other clustering algorithm.

To test the stability and reliability of the clusters, a random hold-out sample of 250 cases was removed from the data base, and the remaining 426 cases were subjected to K-Means clustering. The resulting 10 clusters created out of the 426 cases closely resembled in those of the full 676-case sample in shape and elevation, and there was less than 10% "migration" of individuals from one cluster to another between the 426 and the 676 cluster solutions.

Validity considerations

As noted above, the 10 ARCS ABE clusters were created solely from the 676 learners' scores on the 11 reading tests listed above; no demographic variables from the Background Questionnaire (BQ) were used in the cluster analyses. This enabled the investigators to use BQ means and crosstabs for each cluster to examine independently whether individuals in each cluster shared other relevant factors, such as linguistic and educational histories, self-reports of reading habits, and self-reports of reading difficulties.

Throughout the 10 clusters, this was found to be the case. For example, the two Beginning reader clusters reported the highest rates of early reading problems in childhood and high rates of previously diagnosed reading disability. As would be expected in a clinical setting, they also evidenced the greatest impairments in short term and working memory on the WAIS Digits Forward and Backward subtests, two variables that were not used in the cluster analysis. In another example, one Low Intermediate Cluster which had GE 2 mean scores in both of the ARCS oral vocabulary measures turned out to be made up of over 95% English Language learners who had very low mean levels of native language education.

Based on these kinds of analyses of BQ data, it was concluded that the 10 ARCS ABE/ASE clusters were not simply artifacts of the particular tests used or the Ward's or K-Means clustering methods. The 10 ARCS clusters appeared to be strongly related to the life experiences and reported literacy histories and behaviors of the 676 participants.

The validity of any cluster solution can be partly inferred by whether the results are reproducible. Muth (2004) used a battery of array of assessments similar to the ARCS, supplemented with some Woodcock-Johnson tests, to create a set of reading profile clusters for inmates in the federal prison system. The resulting clusters were very similar to the ARCS, with the exception that in Muth's study a higher proportion of inmates were placed in beginning reading clusters. Strucker, Yamamoto, and Kirsch (2007) used a shorter list of assessments[3] to test a convenience sample of 1084 ABE/ASE and ESL enrollees. Latent class analysis of ABE and ESL enrollees together produced a different array of clusters from the ARCS, which used separate ABE and ESL databases. However, the latent classes containing native speakers and advanced ESL learners resembled the similar ARCS ABE clusters in term of their patterns of strengths and needs. In addition, the relationship of that study's BQ variables to reading data was very similar to that observed in the ARCS. A recent Canadian study, (Grenier, et al., 2008) found similar results in a latent class analysis of the reading components skills of Francophone and Anglophone adults with limited literacy skills.

Ultimately, the validity of "instructionally relevant" clusters such as those in the ARCS can only be determined by studies that focus on educational outcomes. The IES/NICHD Adult Literacy Network studies that are being concluded as of this writing may be able to contribute to this question, because several of those studies have attempted to target instructional interventions to learner profiles that are based on reading components scores.

[3] This list was based on factor analysis of the ARCS data.

ARCS and the Assessment Strategies and Reading Profiles (ASRP) Website

As discussed earlier, the ARCS administered many more tests than ABE or ESL programs would ever have the time or the inclination to administer in the real world. A key finding of the ARCS was that a much shorter list of tests could produce essentially the same constellation of ABE clusters as the longer ARCS research battery. A stepwise regression revealed that nearly all of the variance in Silent Reading Comprehension in the ARCS ABE/ASE sample could be accounted by the following three tests: DAR Oral Reading, DAR Word Recognition, DAR Spelling, and DAR Word Meaning.[4] A later factor analysis of all ARCS test results conducted by Kentaro Yamamoto of ETS also confirmed the heavy contribution of the above four tests.

As a result, when Rosalind Davidson designed the Match-A-Profile feature for the ASRP website, there was substantial empirical and theoretical justification for using only the same tests: word recognition, word meaning, oral reading rate, and spelling – in addition to silent comprehension - to create the ASRP profiles. That means that when ABE teachers use the ASRP site to enter a learners' scores for those four tests in the Match-A-Profile, they can be reasonably confident that that they can match their learners to the appropriate ASRP profiles, and that the ASRP profiles in turn embody the range of ABE profiles described in the ARCS.

Limitations of the ARCS

Because the ARCS used a convenience sample of ABE/ASE learners from only seven participating states and the learners were assessed and interviewed in a particular historical moment, the percentages of learners in each of the 10 ARCS ABE/ASE clusters should not be viewed as an accurate estimate of the actual percentages of learners in those clusters in the ABE system today or in 1999-2000. Most notably, because the ARCS did not attempt assess learners in volunteer programs[5] or in prisons, beginning readers were probably under-represented in the sample. In addition, beginners may have been under-represented because the ARCS did not sample learners in several southern states known to have the highest percentages of adult beginning readers as revealed by the 1992 National Adult Literacy Survey (Kirsch, et al., 1993) and confirmed by the 2003 National Assessment of Adult Literacy (National Center for Education Statistics, 2005). And, as discussed earlier, beginning ESL learners who spoke languages other than Spanish were not included in the ESOL part of the study.

[4] These results are in keeping with Gough and Hillinger's "simple view of reading" (1980), the work of Charles Perfetti and his colleagues (1985), and later causal modeling by Carver (2001) on the relationship of print and meaning components to reading comprehension.

[5] While the ARCS was in the field, several participating programs confided that they had recently begun referring beginning readers to local volunteer programs, because beginners tend to make very slow measurable progress, and it was more difficult for programs to earn performance-based payments for them.

References

Carver, R. & David, A.H. (2001). Investigating reading achievement using a causal model. *Scientific Studies of Reading*. Volume 5, Issue 2 April 2001, pages 107 – 140.

Chall, J.S. & Curtis, M.E. (1990). Diagnostic achievement testing in reading. In Reynolds and Kamphaus (Eds.), *Handbook of psychological and educational assessment of children, V. I.* New York: Guilford Press.

Davidson, R.K., Strucker, J. (2002). Patterns of word-recognition errors among adult basic education native and nonnative speakers of English. *Scientific studies of reading, 6, 3, 299-315.* New Jersey: Erlbaum.

Gough, P. B., & Hillinger, M. L. (1980). Learning to read: An unnatural act. *Bulletin of the Orton Dyslexia Society, 30*, 179-196.

Grenier, S., Jones, S., Strucker, J., Murray, T.S., Gervais, G. & Brink, S.(2008). *Learning Literacy in Canada: Evidence from the International Survey of Reading Skills.* Ottawa, CA: Statistics Canada. Accessed as a pdf 10-30-08: http://www.statcan.ca/english/research/89-552-MIE/89-552-MIE2008019.htm

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult illiteracy in America: A first look at the results of the National Adult Literacy Survey.* Washington, D.C.: National Center for Education Statistics, U.S. Department of Education.

Lorr, M. (1983). *Cluster analysis for the social sciences.* San Francisco: Jossey-Bass.

Muth, W.R. (2004). *Performance and perspective: Two assessments of federal prisoners in literacy programs.* Unpublished doctoral dissertation, George Mason University.

National Center for Education Statistics. (2005). *Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy.* Author: Washington, DC: http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007480

Perfetti, C.A. (1985). *Reading ability.* New York: Oxford University Press.

Rosner, J. (1975) *Helping children overcome learning disabilities.* New York: Walker.

Strucker, J. (1995). *Patterns of reading in adult basic education.* Unpublished doctoral dissertation, Cambridge, MA: Harvard University Graduate School of Education, Gutman Library.

Strucker, J., Davidson, R., & Reddy, L. (2001). *ARCS interviewer training manual.* Cambridge, MA: NCSALL. Available as a pdf on the ASRP website: include URL here once we have it.

Strucker, J. & Davidson, R. (2003). The Adult reading components study. NCSALL research brief. http://www.ncsall.net/fileadmin/resources/research/brief_strucker2.pdf

Strucker, J., Yamamoto, K., & Kirsch, I. (2007). *The relationship of the component skills of reading to IALS performance:tipping points and five classes of adult literacy learners.* NCSALL Report #29 (March 2007): http://www.ncsall.net/?id=29

Assessments

ABLE. (1986). *Adult basic learning exam.* San Antonio, TX: Pearson.

Dunn, L.M., Dunn, L.M., & Dunn, D.M. (1997). *Peabody Picture Vocabulary Test III (PPVT-III).* Circle Pines, MN: AGS.

CASAS. (2000). *Comprehensive adult student assessment systems.* SanDiego, CA: CASAS.

Language Assessment Battery-English (LAB). (1982). New York: New York City Board of Education.

Rapid Automatized Naming. Tests supplied by and used in the ARCS with permission of Professor Marianne Wolf, Tufts University.

Rosner, J. (1975). Test of Auditory Awareness Skills. In *Helping children overcome learning disabilities.* New York: Walker.

Roswell, F.G.& Chall, J.S. (1992). *Diagnostic Assessments of Reading (DAR)*. Itaska, IL: Riverside.

Dunn, L.M., Lugo, D.E., Eligio, R., & Dunn, L. (1986). *Test de Vocabulario en Imagenes Peabody (TVIP).* Bloomington, MN: Pearson Assessments.

TABE. (1987) *Test of adult basic education.* Monterey, CA: CTB/McGraw-Hill.

Woodcock, R.W. & Muñoz, A.F. (1987). *Bateria Woodcock-Muñoz.* Circle Pines, MN: AGS.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Test-Revised (WRMT-R).* Circle Pines, MN: AGS.

Appendix I: Nonnative speakers of English in the ARCS

We knew we would be able to recruit interviewers who could interview and test in Spanish speaking ESOL students. But the covering range of other native languages present in ESOL classes in the US proved to be quite daunting. According to the 1992

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 difficult 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.[6] In practice, this usually meant students in "intermediate" and above ESOL classes. To help us decide whom we could test, we consulted teachers and administrators at the actual program 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/levels of their students they thought 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 the core of English language reading tests, but additional tests would be given based on their appropriateness. Enrollees in all levels of ABE (including ASE and GED) and enrollees 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 additionally tested in Spanish reading using parts of the Woodcock-Munoz Battery and in Spanish vocabulary with the Test de Vocabulario en Imagenes Peabody (TVIP). Anyone, regardless of native language, 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. However, ESOL students who were not native speakers of Spanish were not tested in these areas because neither translations nor scoring would have been possible in their native languages. 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.

[6] As a reminder, it should be noted that these limitations in the ESOL enrollee sample did not affect the ABE enrollee sample, the sample upon which ASRP website profiles were based.

Appendix II: Table of five ARCS testing protocols

Test names See References for complete test information	**Native Spanish Speakers**			**Native Speakers of Other Languages**	**All other ABE/ASE enrollees (except native Spanish speakers)**
	Beg/Int ESL	Adv ESL	ABE/ ASE	Adv ESL	
ARCS Questionnaire	X*	X	X	X	X
Diagnostic Assessments of Reading [1]	X	X	X	X	X
***Wechsler Adult Intelligence Scale-III* (Digit Span)**	X*	X*	X & X*	X	X
***Wechsler Adult Intelligence Scale-III* (Information)**	X*	X	X	X	X
Rapid Automatized Naming [2]	X*	X*	X & X*		X
Test of Auditory Analysis Skills	X*	X*	X & X*		X
Peabody Picture Vocabulary Test	X	X	X	X	X
Woodcock Reading Mastery [3]	X	X	X	X	X
Language Assessment Battery	X	X	X	X	X[5]
Test de Vocabulario en Imagenes Peabody	X	X	X		
Woodcock-Muñoz Bateria [4]	X	X	X		

****in Spanish***

[1] ***Diagnostic Assessments of Reading***; [Word Recognition; Word Analysis (consonant sounds); Oral Reading (fluency; rate); Word Meaning; Silent Reading Comprehension; Spelling]

[2] ***Rapid Automatized Naming*** [letters, numbers, objects, colors, letters/numbers]
[3]***Woodcock Reading Mastery*** [Word Attack]
[4]***Woodcock-Muñoz Bateria*** [Identificación de letras y palabras; Análisis de palabras; Comprensión de textos]
[5] for native speakers of other than English or Spanish