



Importance of Learner Characteristics in Intelligent Tutoring for Adult Literacy

John Hollander, John Sabatini, Art Graesser, Daphne Greenberg, Tenaha O'Reilly & Jan Frijters

To cite this article: John Hollander, John Sabatini, Art Graesser, Daphne Greenberg, Tenaha O'Reilly & Jan Frijters (2023): Importance of Learner Characteristics in Intelligent Tutoring for Adult Literacy, *Discourse Processes*, DOI: [10.1080/0163853X.2023.2203543](https://doi.org/10.1080/0163853X.2023.2203543)

To link to this article: <https://doi.org/10.1080/0163853X.2023.2203543>



Published online: 17 May 2023.



[Submit your article to this journal](#)




[View related articles](#)



[View Crossmark data](#)



Importance of Learner Characteristics in Intelligent Tutoring for Adult Literacy

John Hollander ^a, John Sabatini^a, Art Graesser^a, Daphne Greenberg^b, Tenaha O'Reilly^c, and Jan Frijters^d

^aInstitute for Intelligent Systems, The University of Memphis; ^bLearning Sciences, Georgia State University; ^cEducational Testing Service, Princeton, NJ; ^dChild and Youth Studies, Brock University

ABSTRACT

Adult literacy learners are characterized by their diversity, both in terms of educational histories and cognitive skill sets. Accounting for the specific strengths and weaknesses of each learner is vital to the assessment of literacy gains and optimization of educational systems. We examined pre- and post-difference scores on a component reading skills assessment battery collected before and after an instructional program that included an adult comprehension-focused intelligent tutoring system. By characterizing learners during instruction, we examined differential gains in foundational reading skills. Most learners made gains in reading skills above the word recognition and decoding level; readers who were classified as “conscientious” (who performed slowly but accurately) tended to make the most substantial gains. We conclude that this hybrid instructional program may be an effective educational environment for adult literacy and describe how characterizing learners via integrating assessments into adaptive instructional practice may improve efficiency and effectiveness.

Introduction

The results of the 2013 Programme for International Assessment of Adult Competencies estimate that 19% of adults worldwide (17.5% in the United States) are classified as at or below Level 1 Literacy Proficiency and another 33% at Level 2 (OECD, 2013). Readers at or below Level 1 may struggle with short texts and literacy tasks that involve single operations, such as constructing the literal understanding of a passage or searching for a piece of information. Readers at Level 2 are comfortable with slightly longer texts and can generally begin to integrate textual information by comparing and contrasting or making low-level inferences. Most adults at Level 2 are considered below postsecondary literacy proficiency, whereas adults at or below Level 1 possess very limited literacy skills. In short, the population of struggling adult readers is substantial and will remain so without advances in adult literacy research and educational technology.

Adult literacy programs seeking to provide educational resources to this population often lack important tangible resources, such as funding, but also less-tangible resources, such as consistent, effective, and theoretically driven assessment and curricular frameworks (Greenberg, 2008). Without these, it can be difficult to understand the needs of individual learners, tailor their educational experiences, and measure their growth over time. Personalized instruction is an especially vital goal in adult education because adult learners are an immensely diverse and complex population. Adult learners come from highly variable socioeconomic backgrounds and have disparate linguistic and educational histories and goals (Tamassia et al., 2007). Compared to adolescent students who read at similar levels, adult

literacy learners are more likely to be non-native English speakers (or to speak nonstandard English), have undiagnosed reading and learning disabilities, and may have developed compensatory cognitive processes or strategies to mitigate a lifetime of reading difficulties (Sabatini, O'Reilly et al., 2019; Washington et al., 2013).

It is imperative to account for considerable diversity in learner characteristics to optimize instruction for adult reading comprehension. This presents a challenge to educators and researchers: How can the characteristics of individuals from such a diverse population be operationalized and measured to provide fast, reliable, and actionable adaptation? One possible approach involves using computerized educational systems that generate large amounts of user interaction data. Response speed and accuracy are readily and abundantly generated data in most computerized educational systems. These types of data are suitable for making inferences about readers since proficient reading relies on quickly and accurately coordinated skills (Feller et al., 2020; C. A. Perfetti, 1985; Perfetti & Adlof, 2012). We sought to address the challenges associated with high learner variability by leveraging what we know about cognitive reading processes to analyze and utilize speed and accuracy data from adult learners in a computerized educational system designed to teach reading skills.

Proficient reading comprehension is an intricate process that involves the use of several continuous, overlapping skills (Graesser & McNamara, 2011; Perfetti & Adlof, 2012). These skills range from the recognition of letter-sound correspondences to the construction of complex situation models of text meaning. Over time and with practice, these skills ideally become automatized, occurring more rapidly and with less effort (Ehri, 2005), which allows more time and effort to be devoted to comprehension and reasoning. While one does not need to fully master lower-level skills in order to perform higher-level ones, readers with severe weakness in foundational reading skills may be forced to expend significant cognitive resources to overcome these obstacles. This depletes the resources available for higher level comprehension and discourse processes, resulting in diminished overall reading comprehension abilities (Cain & Barnes, 2017; Cain & Oakhill, 2012; Wang et al., 2019). Struggling readers who lack the ability to rapidly integrate new textual information with their prior knowledge are often unable to meet the demands of complex comprehension tasks. This not only affects their performance on these tasks, but also constrains their ability to learn from texts (McNamara & Magliano, 2009).

However, despite the multifaceted nature of the construct, reading comprehension is typically assessed as a single construct in most educational settings, especially in adult literacy programs, where limited resources often preclude the use of frequent and time-consuming assessment batteries. Even when assessments are available, it is not always clear whether or how this information is used to personalize educational opportunities for learners to maximize growth (Belzer & Greenberg, 2020). Ignoring the component skills involved in reading leaves instruction overgeneralized and suboptimal. One approach to avoiding overgeneralization while accommodating the limited resources of adult education programs is to fine tune student learning through *individualized, efficient assessment and intelligent tutoring systems*.

The objective of this work is to examine how component reading skill assessments and instructional program data may be leveraged to better understand individual reading skill sets and improve educational reading comprehension outcomes for adult literacy learners. More specifically, we examine data from one assessment and one instructional system used as a part of an extensive adult literacy education program (Graesser et al., 2016, 2019). By analyzing within-subjects data from pre- and postinstruction assessments at multiple levels of reading skill and comprehension, as well as data from a computerized, adaptive learning system used during instruction, the characteristics of individual learners can be derived and examined. These information-rich learner models can be used to determine not only *whether* specific instructional practices are effective, but also *for whom* they are most effective and *for what* particular reading skills are impacted.

Reading Inventory and Scholastic Evaluation framework for assessing reading and comprehension

The Reading Inventory and Scholastic Evaluation (RISE) assessment is a subset of the operational reading component skills assessment system known as the Study Aid and Reading Assessment (Sabatini, Weeks et al., 2019). The conceptual framework for the RISE is primarily founded in Perfetti's verbal efficiency theory; proficient readers are able to perform low-level reading skills both accurately and efficiently, reserving cognitive resources for higher-level comprehension and inference generation processes (Feller et al., 2020; C. A. Perfetti, 2001; Sabatini, O'Reilly et al., 2019). The skills assessed include *decoding/word recognition, vocabulary, morphology, sentence processing, reading efficiency, and overall reading comprehension*. Foundational reading skills may interact to contribute to reading comprehension in a complex, nonhierarchical stream; however, performance on each of the RISE subtests yield significant, unique associations with performance on the reading comprehension subtest (Sabatini, Weeks et al., 2019). Further, a lower-skill reader may rely more on context and compensatory behaviors than a highly skilled reader. This is especially true of adult learners, who often develop a wide array of compensatory strategies over the years that are not always optimal to continued growth toward proficiency. The RISE may provide a useful glimpse into the nuances of individual reading skill differences and, in the case of this study, how they may be associated with changes in comprehension as a result of an intervention.

This assessment battery has been the subject of field studies and psychometric validation studies in schools across the United States (Sabatini, Weeks et al., 2019). While original field studies were conducted in grade schools, researchers have recently begun to utilize the RISE in studies of postsecondary students (Feller et al., 2020) and adult literacy learners (Chen et al., 2021; Hollander et al., 2022). Its relative ease of administration allows for low-demand monitoring of component reading skills. The RISE is also vertically scaled, meaning one can make direct comparisons between samples and populations who complete its various forms or item sets (see Sabatini, Weeks et al., 2019 for a technical report).

AutoTutor framework for reading comprehension instruction

AutoTutor is an intelligent tutoring system that uses conversational agents to promote learning in a variety of domains (Graesser, 2016; Nye et al., 2014), including a content package designed specifically for adult literacy instruction. In this system, two computer agents (i.e., a "tutor" agent and a "peer" agent) guide learners using conversation-based dialog interactions while completing sets of tasks selected by human instructors. The versions of AutoTutor that have been tested contain 30 lesson topics that were designed for adult literacy learners, ranging in scope from word-level learning to passage-based reading activities. Many lessons contain contingent branching structures in which questions and texts start at a medium level of difficulty but branch harder or easier questions and texts depending on learners' performance. Most lessons contain between 10 and 35 items, and usually take between 10 and 30 minutes to complete. [Figure 1](#) displays the interface of a sample item. In this example, sentences in a passage are highlighted, and participants are asked to determine whether the highlighted sentences describe a character or an event. Items are often multiple choice, but some items include text input. All items are accompanied by audio output from the agents that guides learners through lessons while adapting to learner performance.

The content of the lessons was designed to align with the U.S. Department of Education's College and Career Readiness Standards for Adult Education (Pimentel, 2013), a multilevel framework of discourse comprehension (Graesser & McNamara, 2011), and a face-to-face reading comprehension curriculum that was successful for low-skill readers in schools (Lovett et al., 2012) and tailored for adults. The multilevel framework outlines five theoretical levels at which comprehension takes place (and can break down) the *surface code*, the *textbase*, the *situation model*, the *genre and rhetorical structure*, and *pragmatic communication*. This provides the theoretical grounding for AutoTutor, with

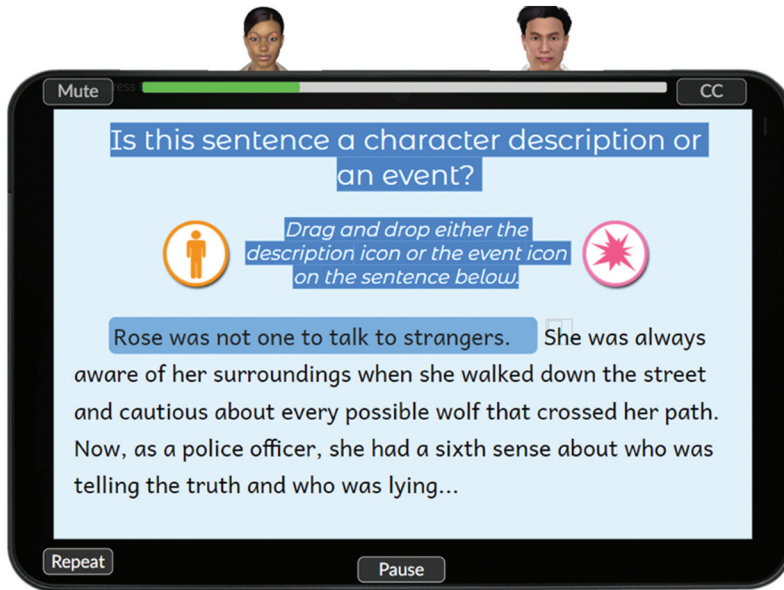


Figure 1. An example AutoTutor item.

a focus largely on higher-level discourse comprehension but some attention to foundational reading skills.

Regarding the curriculum, AutoTutor was designed to emulate PACES, a reading curriculum which has successfully promoted reading gains in remedial-level high school students (Lovett et al., 2012). Accordingly, each lesson targets one of the following: Predicting topic and writer's purpose with text signals and key information, Acquiring new vocabulary with context clues, Clarifying common sources of confusion about the text with clarifying questions, Evaluating, elaborating, and explaining through questioning, or Summarizing, identifying and constructing text structures. A major advantage of this approach is its wide overall construct coverage, so most students are able to take lessons in the domains in which they most need help. An additional advantage of this version of AutoTutor is that it was specifically designed to suit adult literacy learners. It does so by simulating digital environments that reflect how literacy skills are required of adults in pragmatic, naturalistic situations. Tasks like these may be of higher interest and value to adult learners, which increases the likelihood of maintaining engagement and acquiring practical skills (Hollander et al., 2021; Leu et al., 2015). AutoTutor has been used in partnerships with adult literacy programs on a small scale and undergoes continuous refinement.

Items in AutoTutor typically consist of multiple-choice answers to conversation-based questions asked by one of the agents. There are typically three response options to each question. If a learner exhibits a misconception by answering a question incorrectly, the agents will typically explain why a response is incorrect, why correct alternatives are correct, and sometimes where to find the information in the text that can be used to arrive at the correct answer (and refute the previous misconception). After this guidance, learners are usually given another attempt at the item, but these secondary attempts are not factored into the scores used for the difficulty branching procedure described above.

Interventions using AutoTutor have predicted pre- to postinstruction learning gains on some measures of global reading comprehension, including the Woodcock-Johnson III and the RAPID (Chen et al., 2021; Fang et al., 2021). However, the effects on *specific reading component skills have yet to be reported*. While AutoTutor and the PACES curriculum generally target higher-level reading skills than those measured by the RISE, there is some explicit overlap between the two frameworks, such as an emphasis on vocabulary. Further, without the measurement of more

foundational reading skills, the question of whether (and how) comprehension gains are made in isolation or in confluence with growth in other foundational reading skills remains unaddressed. A key aim of this article is to explore how foundational reading component skills are affected by adult literacy-oriented instruction. In this case, we use data collected from AutoTutor as a means to track and infer performance in a hybrid instructional program that includes both human and computer training.

Materials and methods

Data were obtained from three waves of an adult literacy intervention program study. Participants consisted of 252 adult literacy program enrollees in the United States and Canada ($M_{age} = 42.4$, $SD = 13.9$, 74.6% women). Approximately 59.5% of the sample identified as Black, 17.5% multiracial, 11.5% White, and 9.5% Asian. The participants took part in 100 hours of an instructional intervention featuring a blended class consisting of sessions with teacher-led decoding, vocabulary, and comprehension instruction in addition to AutoTutor. Regarding comprehension training, the instructional program primarily followed the PACES curriculum, with AutoTutor lessons corresponding to the in-person lesson topic that day.

Participants completed one form of the RISE before the 100-hour intervention and another form on completion. Each of the five subtests that target component skills typically takes less than 10 minutes to complete for most struggling readers; the general comprehension subtest typically takes 15–20 minutes to complete. To more precisely convey the foundational reading skills targeted by the RISE, brief descriptions and example items are provided below, followed by their observed reliability in typically developing samples across grades 5–12 (Sabatini, Weeks et al., 2019).

Word recognition and decoding

Participants were presented with a word or pseudoword and decided if it (1) is a real word (e.g., *shoulder*), (2) is not a real word (e.g., *plammity*), or (3) sounds exactly like a real word even if it is not spelled correctly (e.g., *brane*). The reliability ranges $\alpha = .89$ –.92 across grade levels.

Vocabulary

Participants see a target word and three options and must decide which option is a synonym of the target (e.g., data: *information*, schedule, star) or a semantic associate of the target (e.g., marine; store, tree, *water*). The reliability ranges $\alpha = .83$ –.90 across grade levels.

Morphology

Participants see a sentence with one word missing and three options, all of which are derivational variants of the same root and choose the word that completes the sentence (e.g., Some scientists believe the moon is an area we'll be able to ____; colonist, colonial, *colonize*). The reliability ranges $\alpha = .87$ –.91 across grade levels.

Sentence processing

Participants see a sentence with one word missing and three options and choose the word that completes the sentence. Response options contain emphasis on relation-signaling words (e.g., *because*, *if*, *although*) as a key element of this subtest because they are particularly diagnostic of struggling readers (e.g., Ana couldn't ride her bike to school ____ the chain fell off.; *because*, instead, although). The reliability ranges $\alpha = .83$ –.87 across grade levels.

Reading efficiency

Also known as the maze task (Shin et al., 2000), participants complete cloze items which are derived from successive sentences selected from a passage. This subtest differs from the previous subtest because it is timed and strings together cohesive, within-passage sequences and target words that can be either functional relation-signals or topical vocabulary. This causes readers to utilize and coordinate several lower-level reading skills at once without involving high-level inference generation (e.g., During the Neolithic Age, humans developed agriculture—what we think of as farming. Agriculture meant that people stayed in one place to grow their *crops/baskets/rings*. They stopped moving from place to place to follow herds of animals or to find new wild plants to *eat/win/cry*. And because they were settling down, people built permanent *shelters/planets/secrets*.). The reliability ranges $\alpha = .92-.95$ across grade levels.

Reading comprehension

This subtest is a condensed version of traditional, passage reading comprehension assessments. This task was designed to assess a reader's surface-level (words and phrases) and gist-based (roughly textbase and situation model) understandings of brief passages, while limiting reliance on deeper conceptual or social reasoning questions. The number of questions that demand domain-specific prior knowledge and deeper conceptual or social reasoning questions are purposefully limited. This subtest contrasts with scenario-based assessments that target higher levels of discourse skills (O'Reilly & Sabatini, 2013; Sabatini et al., 2020). Rather, the RISE comprehension subtest requires participants to locate information, paraphrase, and make low-level inferences at the textbase level. The reliability ranges $\alpha = .60-.83$ across grade levels.

Data processing

We conducted a hierarchical clustering analysis on AutoTutor performance data in order to investigate how instructional and assessment data can be leveraged to identify vital learner characteristics in adult literacy learner populations. Previous studies have conducted clustering analyses on the same interventions' participants and data (Fang et al. (2018, 2021) so we sought to replicate and test the robustness of these findings by conducting similar analyses with stricter exclusion criteria and data processing parameters. We then sought to test differences between RISE scores by cluster to determine which types of learners benefitted from the instructional program on each foundational reading skill. It should be noted that the Fang et al. publications *did not report how the multiple measures of RISE were related to the clusters of learners manifested in AutoTutor performance*.

In preparation for the clustering analyses, we included only responses to the lesson segment prior to the branching of students to more difficult or easier texts and items. This segment always included items that spanned a wide range of difficulty; that is, no floor or ceiling effects were observed at a group level for any lesson. After the branch point in a lesson, some students were routed to a more challenging set of items, while others were routed to a relatively easier set, thus reducing the sample of students who took all items. By limiting the item pool to the first segment, we maximized sample size for the cluster analysis. Next, individual "ceiling items" that more than 98% of participants answered correctly on their first try were removed in order to ensure adequate variance in item accuracy (as opposed to a 100% threshold used by Fang et al., 2018, 2021). This excluded 11.1% of data. Outliers from the remaining AutoTutor response time data (more than three times the interquartile range below the first quartile and above the third quartile) were removed, excluding 3.4% of the data (Fang et al., 2018, 2021). Response times of less than 3 seconds were Winsorized and adjusted to 3 seconds (Fang et al., 2018, 2021). After these criteria were imposed, there were no outliers on either accuracy or time at the aggregate participant or lesson level. No participant's individual distribution of accuracy or time yielded a skewness value below -1 or above 1 . Despite the changes to the processing parameters, all participants were classified identically as in previous studies (Fang et al., 2021). Approximately 38,800 observations remained in the final data set.

Finally, to examine the potential of AutoTutor as an instructional tool for making gains in foundational reading skills, we conducted a series of comparative analyses on pre- and postintervention difference scores on the RISE.

Results

Clustering of readers according to AutoTutor performance

Hierarchical clustering analyses were performed on the AutoTutor accuracy and latency data using Ward's method (Ward, 1963). These analyses were conducted using the R package *cValid* (Brock et al., 2008). We attempted to replicate similar analyses by Fang et al. (2018, 2021), in which a four-cluster solution yielded optimal results. Despite some differences in the data processing parameters (e.g., more conservative exclusion criteria and “ceiling item” accuracy thresholds), our analysis labeled participants identically to the analyses conducted by Fang et al. (2018, 2021). This suggests a successful, generalizable, and convergently robust methodological reproduction.

Based on patterns in accuracy and response speed, Fang et al. (2018, 2021) (see also Chen et al., 2021) labeled clusters as *proficient readers* (accurate and fast, $N = 97$), *underengaged readers* (medium accurate but fast, $N = 93$), *conscientious readers* (medium to accurate but slow, $N = 31$), and *struggling readers* (very inaccurate and slow, $N = 31$). For consistency in reporting, we also use these labels, but with several clarifications. First, the term “proficient” is relative to the specific sample of struggling adult readers who did not read at the eighth grade level—not an absolute standard of proficiency in reading. Further, the “conscientious” and “underengaged” labels imply knowledge of the participants' individual engagement levels, which is only indirectly known. We might say that by responding relatively slow but accurate, it is *as if* the students are behaving conscientiously, while by responding quickly and inaccurately, it is *as if* students were underengaged.

ANOVAs of average item accuracy and response time yield group differences (accuracy: $F(3, 249) = 50.92, p < .001, \eta^2 = .398$; time: $F(3, 249) = 45.92, p < .001, \eta^2 = .356$). Regarding accuracy, post-hoc independent *t*-tests with Bonferroni corrections indicate that the proficient cluster was significantly more accurate than all other clusters (all $p < .001$) and that the struggling cluster was significantly less accurate than all other clusters (all $p < .001$), but the underengaged and conscientious clusters were not significantly different ($p = .56$). Regarding response time, similar analyses indicate that the proficient and underengaged clusters were significantly faster than the conscientious and struggling clusters (all $p < .001$) but not significantly different from each other, and the conscientious cluster was significantly slower than all other clusters (against proficient and underengaged $p < .001$, against struggling $p < .01$). Figure 2 graphically represents these comparisons.

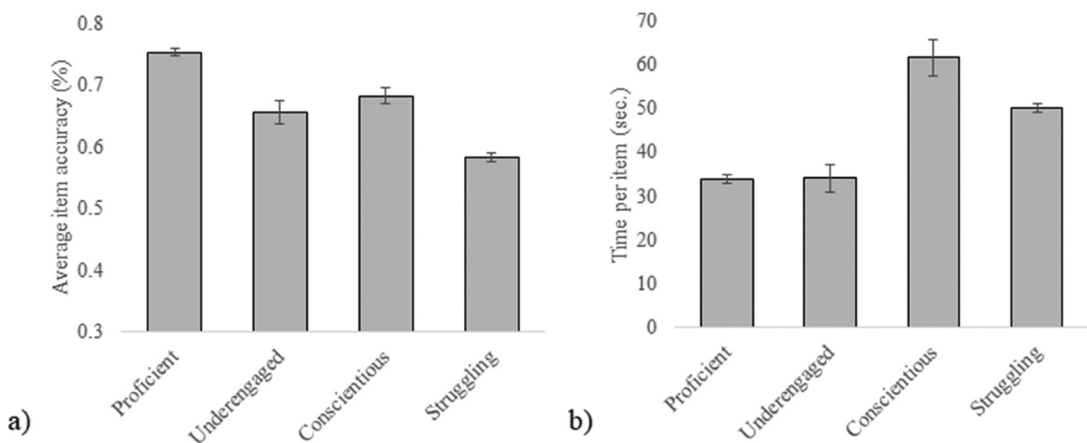


Figure 2. Average accuracy (a) and response time (b) on AutoTutor items as a function of cluster.

Preintervention scores

Our characterization of these groups so far is limited to their general accuracy and speed in high-level reading tasks. However, in order to better understand these groups and provide convergent validity to our characterizations, we conducted a series of six ANOVAs comparing the four clusters on each of the RISE subtests, preintervention. Because the construct coverage of the RISE involves foundational reading skills *and* their efficient integration, we can further observe how our cluster classifications based on speed and accuracy align with the requisite components of proficient reading. Post-hoc comparisons were conducted using Bonferroni alpha-level corrections.

Each of the ANOVAs was statistically significant; marginal means are displayed in Table 1. For word recognition and decoding ($F(3, 178) = 10.42, p < .001, \eta^2 = .15$), vocabulary ($F(3, 177) = 12.91, p < .001, \eta^2 = .18$), and morphology ($F(3, 176) = 8.07, p < .001, \eta^2 = .12$), the proficient cluster scored significantly higher than all other clusters (all $p < .01$). For sentence processing ($F(3, 177) = 6.29, p < .001, \eta^2 = .10$), the proficient cluster scored significantly higher than the underengaged and struggling clusters (both $p < .01$). For efficiency ($F(3, 178) = 25.77, p < .001, \eta^2 = .30$), the proficient cluster scored significantly higher than all other clusters (all $p < .01$), and the underengaged cluster scored significantly higher than the conscientious and struggling cluster (both $p < .01$). For reading comprehension ($F(3, 178) = 12.31, p < .001, \eta^2 = .17$), the proficient cluster scored significantly higher than the underengaged and struggling clusters (both $p < .001$).

Reading component skill gains

To determine AutoTutor's potential as a tool for tracking the growth of reading skills, we analyzed the post- and predifference scores on each RISE subtest by cluster. Within-subtest pre- and postscores were generally highly correlated ($r_{\text{WRDC}} = .71, r_{\text{VOC}} = .79, r_{\text{MORPH}} = .78, r_{\text{SEN}} = .58, r_{\text{EFFIC}} = .80, r_{\text{RC}} = .62$, all $p < .001$). There were no outliers in the overall distribution of difference scores for each subtest (no values more extreme than 1.5 times the interquartile range from the median), and these distributions were all normally distributed (no skewness values less than -1 or more than 1).

To determine whether participants made significant gains in component reading skills, we conducted a repeated-measures ANOVA for each RISE subtest. For each model, cluster was entered as a between-subjects factor, time (pre- and postprogram) was entered as a within-subjects factor, and each model included the cluster-by-time interaction term. To correct for alpha-level inflation, these tests were conducted using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to adjust observed p values, and critical p values were adjusted at a false discovery rate of .05. The main effects of cluster and time were both significant in each model, except for word recognition and decoding, for which only cluster was significant. The significant main effect of time indicates that participants overall achieved higher scores on the postassessment than the preassessment, indicating an increase in literacy skills, except for word recognition and decoding. The significant

Table 1. Descriptive Statistics of Preinstruction Component Skills by Cluster

| Cluster | WRDC | VOC | MORPH | SEN | EFFIC | RC |
|--------------------|---------------|----------------|----------------|----------------|----------------|----------------|
| <i>Preprogram</i> | | | | | | |
| Proficient | 247.56 (8.61) | 256.49 (11.48) | 250.79 (8.48) | 251.31 (5.87) | 249.61 (8.15) | 259.41 (7.74) |
| Underengaged | 242.37 (8.02) | 249.07 (10.22) | 246.62 (8.43) | 247.74 (7.23) | 243.44 (7.34) | 253.31 (7.30) |
| Conscientious | 240.53 (7.26) | 248.05 (9.22) | 243.63 (9.58) | 246.95 (6.71) | 237.37 (6.12) | 255.32 (8.45) |
| Struggling | 238.52 (6.24) | 242.70 (7.31) | 241.91 (8.66) | 245.57 (6.95) | 237.04 (5.28) | 250.43 (4.96) |
| <i>Postprogram</i> | | | | | | |
| Proficient | 248.06 (8.57) | 260.27* (9.74) | 252.48 (7.60) | 251.69 (6.79) | 251.92* (7.94) | 262.00* (6.98) |
| Underengaged | 242.12 (7.57) | 251.6 (11.47) | 248.56* (8.59) | 249.88 (6.11) | 245.04 (7.93) | 255.83* (7.58) |
| Conscientious | 240.5 (7.21) | 253.13* (9.53) | 247.80* (8.87) | 250.33* (6.55) | 240.33 (8.00) | 257.27 (9.75) |
| Struggling | 238.48 (7.01) | 243.91 (9.38) | 240.74 (8.51) | 247.13 (6.58) | 238.17 (7.44) | 249.73 (6.02) |

*Significant pre- and postcontrasts within-group/subtest combination, $p < .05$ with Bonferroni correction applied.

main effect of cluster indicates some significant pairwise group differences between clusters in each reading skill. The cluster-by-time interaction was not statistically significant in any of the models. These effects are likely impacted by uneven group size and low statistical power to observe significant interaction terms. In order to examine the hypothesis that the groups may have made differential learning gains in each subtest, we also included pre- and postcontrasts for each group with Bonferroni corrections. These contrasts indicated that the proficient group made significant gains in vocabulary (Cohen's $d = .36$), efficiency ($d = .29$), and reading comprehension ($d = .35$); the underengaged group made significant gains in morphology ($d = .23$) and reading comprehension ($d = .34$); and the conscientious group made significant gains in vocabulary ($d = .54$), morphology ($d = .45$), and sentence processing ($d = .51$); contrasts were not significant for the struggling group. The results of each model are displayed in Table 2. The marginal means for pre- and postprogram scores by cluster on each subtest is displayed in Table 1. In addition to marginal means, Table 1 denotes which of these post-hoc contrasts were statistically significant. Figure 3 illustrates the pattern of results in more detail.

Table 2. Results of Repeated-Measures ANOVAs for All Reading Component Subtests

| | WRDC | | VOC | | MORPH | |
|----------------|-----------|---------------|-----------|---------------|-----------|---------------|
| | <i>df</i> | <i>F</i> | <i>df</i> | <i>F</i> | <i>df</i> | <i>F</i> |
| Intercept | 1 | 129,644.15*** | 1 | 75,752.86*** | 1 | 104,506.41*** |
| Cluster | 3 | 13.57*** | 3 | 16.525*** | 3 | 9.63*** |
| Time | 1 | .19 | 1 | 11.76*** | 1 | 6.73** |
| Time × cluster | 3 | .03 | 3 | .737 | 3 | 1.86 |
| | SEN | | EFFIC | | RC | |
| | <i>df</i> | <i>F</i> | <i>df</i> | <i>F</i> | <i>df</i> | <i>F</i> |
| Intercept | 1 | 204,950.97*** | 1 | 136,079.29*** | 1 | 173,597.31*** |
| Cluster | 3 | 5.19** | 3 | 28.44*** | 3 | 19.18*** |
| Time | 1 | 10.87*** | 1 | 7.95** | 1 | 3.92* |
| Time × cluster | 3 | 1.71 | 3 | .416 | 3 | 1.18 |

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

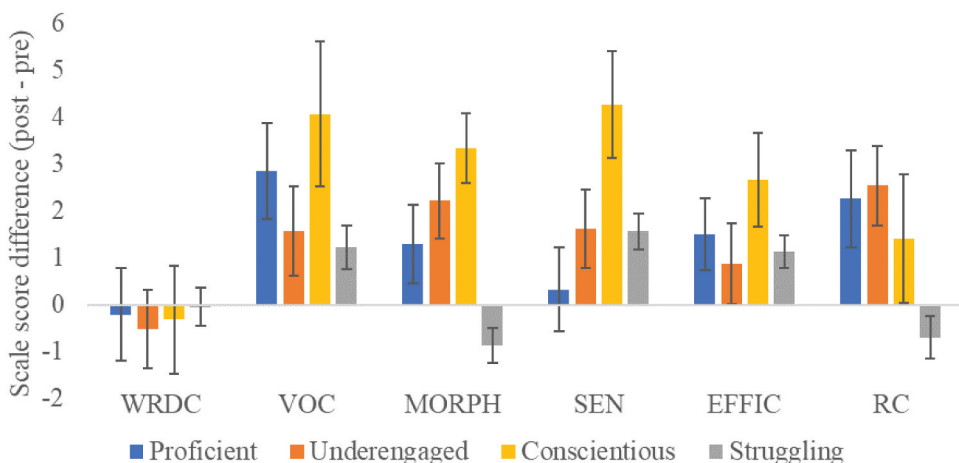


Figure 3. Difference scores for reading component skills by cluster. Error bars represent standard error of the mean.

Discussion

Adult literacy rates remain a significant concern worldwide despite advances in educational research and technology (Greenberg, 2008; OECD, 2013). Adults with low literacy are an incredibly diverse population in several respects, perhaps most vitally in the myriad of cognitive component skill weaknesses that can present obstacles to proficient reading. The results of this study underscore the importance of accounting for learner characteristics when designing assessments and interventions for adult literacy learners. Using AutoTutor, an adult literacy-focused learning environment designed around a generally high-level discourse processing framework, we were able to categorize learners by their accuracy and timing data during instructional tasks and observe how these performance characteristics may be indicative of strengths and weaknesses in foundational reading skills. Further, we analyzed learning gains in component reading skills based on these learner classifications. By doing so, we can observe not only whether the instruction is effective, but also for whom it is most effective.

The results of the clustering analysis demonstrate one method of describing and accounting for diversity in learner characteristics. By describing the clusters based on their speed and accuracy during AutoTutor lessons, we can begin to characterize the experience of learners in the overall hybrid instructional program while gaining some insight into their reading ability. While speed and accuracy are continuous measures, we generally observed a 2×2 factorization in these data. Decades of research demonstrate that proficient reading is both fast and accurate (Feller et al., 2020; C. A. Perfetti, 1985; Perfetti & Adlof, 2012). When readers are fast but inaccurate, one possible explanation is that they may be experiencing fluctuations of engagement. While there may be other explanations, reading speed is typically coupled with sustained, accurate word reading (Ehri, 2005; Perfetti & Adlof, 2012). If participants who perform quickly but inaccurately have inadequate foundational skill to support rapid, accurate word and text processing, then they would need to slow down and engage additional time and effort to ensure accurate responding. Conversely, when readers are moderately accurate but slow, one explanation could be that their component reading skills are perhaps weak or not tightly integrated but adequate for attempting tasks as posed in the lessons. As a result, accurate responding requires additional time and effort and may also be accompanied by using compensatory skills, strategies, knowledge, or metacognitive reasoning, all of which would require additional time during tasks. Lastly, if a reader is neither fast nor accurate, then they are neither processing text efficiently nor able to compensate with other skills or strategies.

These characterizations are contextualized and supported by the preprogram RISE data. For example, the proficient reading cluster generally outscored almost every other cluster in almost every subtest, demonstrating that their foundational reading skills are both effective and efficient. Perhaps most interestingly, the conscientious cluster scored lower than the proficient cluster on most of the foundational skill subtests, especially the reading efficiency subtest, but there was no significant difference between these groups in scores on overall reading comprehension subtest. This supports the descriptive characterizations of these clusters. These results also demonstrate how relative weaknesses in foundational skills may be compensated for with the expenditure of time and effort, likely including the use of top-down strategies and metacognitive reasoning, especially in adult learners (Sabatini, O'Reilly et al., 2019).

The mixture of efficient and effective foundational skills, complemented with compensatory strategies and reasoning is consistent with the learning gains data. We generally observed a pattern of positive learning gains for most component processes for the entire group, with the notable exception of word recognition and decoding. The overall improvement of RISE scores may indicate that the instructional program as a whole was effective in improving not just reading comprehension, but also some of the foundational skills that afford readers the cognitive resources to use higher-level discourse processes. These findings are consistent theoretically both with the multilevel discourse processing framework of AutoTutor and the verbal efficiency framework of the RISE.

While the cluster-by-time interaction term was not significant in our model, post-hoc contrasts indicated that the patterns of skills that significantly improved may differ by cluster. For example, according to these contrasts, proficient readers made significant gains in vocabulary, efficiency, and reading comprehension, while conscientious readers made gains in vocabulary, morphology, and

sentence processing. These differences in skill gain patterns may be clues to the cognitive and meta-cognitive mechanisms associated with reading skill gains. Future research should investigate how, when, and for whom each of these reading skills may be most effectively enhanced by instruction. To illustrate, conscientious readers, who are characterized by somewhat accurate but slow responding, may be more dependent on reading strategies for relative weakness in foundational skill efficiency, which may be more malleable to the designed curriculum as it relates to morphology and sentence processing (Lovett et al., 2012). Conversely, the struggling cluster, who are characterized by both slow and inaccurate responding, exhibited minimal if any gains in some of the foundational skill subtests or reading comprehension scores. It could be that the lesson units in AutoTutor do not adequately address the extent of the weaknesses in some low-level skills (or undiagnosed impairments) that may be present in struggling readers. This finding illustrates the need for instruction to be personalized to individual differences; if a learner is not equipped to engage with the instructional material at its intended level of discourse processing, they will likely not make learning gains and should be directed to instructional material that addresses their specific needs. This recommendation is, of course, reliant on effective and efficient assessment of a learner's abilities and diagnostic of areas for improvement.

A few limitations of this work must be addressed in order to properly contextualize the results and interpretations. First, we used AutoTutor as a way to measure performance and progress in the instructional program. However, the program itself was hybrid in nature, including a face-to-face component. As a result, pre- and postreading skills gains cannot be directly attributed to any specific aspect of the hybrid intervention. Rather, AutoTutor is just one element of the instructional program that provides a substantial amount of data. A second limitation is that our description and characterizations of the clusters make some assumptions about their constituent learners. For example, the term “underengaged” assumes that the configuration of that cluster's accuracy and timing data is due to motivation and engagement, and while we may say that this group of participants are responding *as though* they are underengaged, these properties are not directly measured. Additionally, these labels apply to data that was generated from a wide range of activities and question-answering behaviors throughout the AutoTutor lessons and do not only apply to reading processes, unlike more tightly controlled reading profile work (such as Hyönä et al., 2002). However, we base these labels on previous literature assessing the patterns of engagement in similar complex response data (Chen et al., 2021) and cautiously utilize them for consistency.

While this research provides some insight into the potential for foundational reading skill assessment in adult literacy education, more research must be done to examine the real-time integration of assessment and instruction. Future work should consider how to more effectively help adult readers with significant weaknesses in foundational reading skills, such as the struggling cluster observed in this work. More broadly, future work should also seek to determine how initial reading skill assessments predict future performance and growth. In this study, clustering classifications were applied post-hoc using intelligent tutoring systems data, but an ideal system may be able to make real-time classifications of the needs and risks of learners based on their preinstructional assessment data, including prior education and English language learning status, and their mid-instructional performance data. Finally, future work should seek to map the frameworks of *comprehension*-focused tools such as AutoTutor (including at the lesson- and item-level) to align with assessment frameworks that span *foundational* and comprehension skills, such as the RISE. This would allow for more efficient, covert assessment, and, as a result, more accurate and effective educational technology interventions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through grants [R305C120001], [R305A190522], and [R305A200413], and the National Science Foundation under the award The Learner Data Institute [award #1934745].

ORCID

John Hollander  <http://orcid.org/0000-0002-3270-7495>

Data availability statement

This study was not pre-registered. Due to the nature of this research, supporting data is not publicly available. Please contact the authors for questions regarding data access.

References

- Belzer, A., Greenberg, D. (2020). The Gordian knot of adult basic education assessment: Untangling the multiple audiences and purposes, and L. Hill (Ed.), *Assessment and evaluation in adult and continuing education* (pp. 57–71). Stylus.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031>
- Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). cValid: An R package for cluster validation. *Journal of Statistical Software*, 25(4), 1–22. <https://doi.org/10.18637/jss.v025.i04>
- Cain, K., & Barnes, M. A. (2017). Reading comprehension. In K. Cain, D. Compton, & R. Parrila (Eds.), *Theories of reading development* (pp. 257–282). John Benjamins.
- Cain, K., Oakhill, J. (2012). Reading comprehension development from 7 to 14 years: Implications for assessment. In J. P. Sabatini & E. R. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences* (pp. 143–175). Rowman and Littleford Education.
- Chen, S., Fang, Y., Shi, G., Sabatini, J., Greenberg, D., Frijters, J., & Graesser, A. C. (2021). Automated disengagement tracking within an intelligent tutoring system. *Frontiers in Artificial Intelligence*, 3. <https://doi.org/10.3389/frai.2020.595627>
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9(2), 167–188. https://doi.org/10.1207/s1532799xssr0902_4
- Fang, Y., Lippert, A., Cai, Z., Chen, S., Frijters, J. C., Greenberg, D., & Graesser, A. C. (2021). Patterns of adults with low literacy skills interacting with an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, 1–26. <https://doi.org/10.1007/s40593-021-00266-y>
- Fang, Y., Shubeck, K., Lippert, A., Cheng, Q., Shi, G., Feng, S., Chen, S., Cai, Z., Pavlik, P., Frijters, J., Greenberg, D., & Graesser, A. (2018). Clustering the learning patterns of adults with low literacy skills interacting with an intelligent tutoring system. *Proceedings of the 11th International Conference on Educational Data Mining*, 348–354.
- Feller, D. P., Magliano, J., Sabatini, J., O'Reilly, T., & Kopatch, R. D. (2020). Relations between component reading skills, inferences, and comprehension performance in community college readers. *Discourse Processes*, 57(5–6), 473–490. <https://doi.org/10.1080/0163853X.2020.1759175>
- Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, 26(1), 124–132. <https://doi.org/10.1007/s40593-015-0086-4>
- Graesser, A. C., Cai, Z., Baer, W. O., Olney, A. M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the center for the study of adult literacy. *Adaptive Educational Technologies for Literacy Instruction*, 288–293. <https://doi.org/10.4324/9781315647500>
- Graesser, A. C., Greenberg, D., Olney, A., & Lovett, M. W. (2019). Educational technologies that support reading comprehension for adults who have low literacy skills. *The Wiley Handbook of Adult Literacy*, 471–493. <https://doi.org/10.1080/02702711.2021.1888360>
- Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371–398. <https://doi.org/10.1111/j.1756-8765.2010.01081.x>
- Greenberg, D. (2008). The challenges facing adult literacy programs. *Community Literacy Journal*, 3(1), 39–54. <https://doi.org/10.25148/CLJ.3.1.009480>
- Hollander, J., Sabatini, J., & Graesser, A. (2021). An intelligent tutoring system for improving adult literacy skills in digital environments. *COABE Journal*, 10(2), 59–64. <https://doi.org/10.31234/osf.io/t9xva>

- Hollander, J., Sabatini, J., & Graesser, A. (2022). How item and learner characteristics matter in intelligent tutoring systems data. M. M. Rodrigo, N. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners' and doctoral consortium. AIED 2022. Lecture notes in computer science* (Vol. 13356, pp. 520–523). Springer. https://doi.org/10.1007/978-3-031-11647-6_106
- Hyönä, J., Lorch, R. F., Jr, & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94(1), 44. <https://doi.org/10.1037/0022-0663.94.1.44>
- Leu, D. J., Kiili, C., & Forzani, E. (2015). Individual differences in the new literacies of online research and comprehension. In *Handbook of individual differences in reading* (pp. 277–290). Routledge. <https://doi.org/10.4324/9780203075562>
- Lovett, M. W., Lacerenza, L., De Palma, M., & Frijters, J. C. (2012). Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities*, 45(2), 151–169. <https://doi.org/10.1177/0022219410371678>
- McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of Learning and Motivation*, 51, 297–384. [https://doi.org/10.1016/S0079-7421\(09\)51009-2](https://doi.org/10.1016/S0079-7421(09)51009-2)
- Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427–469. <https://doi.org/10.1007/s40593-014-0029-5>
- OECD. (2013). *OECD skills outlook 2013: First results from the survey of adult skills*. <https://doi.org/10.1787/9789264204256-en>
- O'Reilly, T., & Sabatini, J. (2013). Reading for understanding: How performance moderators and scenarios impact assessment design. *ETS Research Report Series*, 2013(2), i–47. <https://doi.org/10.1002/j.2333-8504.2013.tb02338.x>
- Perfetti, C., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. *Measuring Up: Advances in How We Assess Reading Ability*, 3–20. <https://doi.org/10.7764/onomazein.42.12>
- Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.
- Perfetti, C. A. (2001). The lexical basis of comprehension skill, and D. S. Gorfien (Ed.), *On the consequences of meaning selection* (pp. 67–86). American Psychological Association.
- Pimentel, S. (2013). *College and career readiness standards for adult education*. Office of Vocational and Adult Education. US Department of Education.
- Sabatini, J., O'Reilly, T., Dreier, K., Wang, Z. (2019). Cognitive processing challenges associated with low literacy adults, and D. Perin (Ed.), *The Wiley handbook of adult literacy* (pp. 15–39). John Wiley & Sons.
- Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2020). Engineering a twenty-first century reading comprehension assessment system utilizing scenario-based assessment techniques. *International Journal of Testing*, 20(1), 1–23. <https://doi.org/10.1080/15305058.2018.1551224>
- Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinberg, J., & Chao, S.-F. (2019). SARA reading components tests, RISE forms: Technical adequacy and test design, 3rd edition. *ETS Research Report Series*, 2019(1), 1–30. <https://doi.org/10.1002/ets2.12269>
- Shin, J., Deno, S. L., Robinson, S. L., & Marston, D. (2000). Predicting classroom achievement from active responding on a computer-based groupware system. *Remedial and Special Education*, 21(1), 53–60. Article 1. <https://doi.org/10.1177/074193250002100107>
- Tamassia, C., Lennon, M., Yamamoto, K., & Kirsch, I. (2007). *Adult education in America: A first look at results from the adult education program and learner surveys*. Princeton, NJ: Educational Testing Service.
- Wang, Z., Sabatini, J., O'reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology*, 111(3), 387. <https://doi.org/10.1037/edu0000302>
- Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>
- Washington, J. A., Terry, N. P., & Seidenberg, M. S. (2013). Language variation and literacy learning: The case of African American English. *Handbook of Language and Literacy: Development and Disorders*, 204–221. <https://doi.org/10.1002/9781119606987>