# Adapting for Scalability: Automating the Video Assessment of Instructional Learning

Amy M. Roberts
*University of Nebraska*

Jennifer LoCasale-Crouch, Bridget K. Hamre, Jordan M. Buckrop
*University of Virginia*

**Abstract**

Although scalable programs, such as online courses, have the potential to reach broad audiences, they may pose challenges to evaluating learners' knowledge and skills. Automated scoring offers a possible solution. In the current paper, we describe the process of creating and testing an automated means of scoring a validated measure of teachers' observational skills, known as the Video Assessment of Instructional Learning (VAIL). Findings show that automated VAIL scores were consistently correlated with scores assigned by the hand scoring system. In addition, the automated VAIL replicated intervention effects found in the hand scoring system. The automated scoring technique appears to offer an efficient and reliable assessment. This study may offer additional insight into how to utilize similar techniques in other large-scale programs and interventions.

*Keywords:* automated assessment, scalability, teacher education

Roberts, A. M., LoCasale-Crouch, J., Hamre, B. K., & Buckrop, J. M. (2017). Adapting for scalability: Automating the video assessment of instructional learning, *Online Learning 21*(1), 257-272. doi: 10.24059/olj.v21i1.961

## Introduction

Implementing large-scale evidence-based programs offers a promising means of reaching broad audiences (Franks & Schroder, 2013). Massively Open Online Courses demonstrate one-way educational content is being disseminated widely (Vale & Littlejohn, 2014). The use of online courses has increased among various groups of professionals, including teachers, to advance learners' technical knowledge and skills (Gill, 2011; U.S. Department of Education,

National Center for Education Statistics, 2016). Although recent shifts towards large-scale programs are promising, they pose challenges to assessment, particularly the assessment of learners' skills. Skill assessments are an integral part of most educational programs and are often used to understand individuals' growth trajectories (Biggs & Tang, 2011). Automated coding systems offer a possible means of assessing learners' skills in larger scale programs (Williamson, Xi, & Bryer, 2012).

Given the increased offering of online professional development to teachers (Gill, 2011; Means, Toyama, Murphy, Bakia, & Jones, 2009), the present study sought to adapt a validated measure of teachers' observational skills of effective teacher-child interactions, the Video Assessment of Instructional Learning, or VAIL (Jamil, Sabol, Hamre, & Pianta, 2015), to be automatically scored, rather than manually hand scored, and thus applicable for large-scale interventions. More specifically, the goals of the study were to determine whether the automated VAIL scoring system related to the previously validated hand scoring system and if it was sensitive to intervention effects in a previous professional development program. To achieve these aims, we first correlated automated scores with hand scores and then compared intervention and control group means to determine if the automated VAIL replicated results previously found using the hand scored VAIL. These results are presented along with a discussion of how automated measures may be useful in other large-scale programs.

## Literature Review

As the availability of online coursework grows, course designers and instructors are faced with the challenge of determining how to accurately and efficiently assess students (Palloff & Pratt, 2008). Assessments provide valuable information regarding motivation and progress, and can be used to provide feedback to both learners and instructors. Assessing learners' knowledge and skills from the beginning to the end of a course can also determine the efficacy of a program and may suggest modifications that need to be made (Boston, 2002). In general, assessment questions may be open-ended (short-answer or essays) or closed-ended (true/false or multiple choice.) Open-ended questions provide more thorough information regarding learners' mastery because they require learners to generate responses, rather than simply identifying correct answers from a prescribed list of options (Foddy, 1993). Although, open-ended questions may provide more useful information, they are more arduous to score which may be difficult in courses with large numbers of enrollees (Landauer, Laham, & Foltz, 2003).

Assessing open-ended items typically relies on the knowledge and expertise of human raters, such as instructors or teaching assistants, to manually score and make personal judgments for each response. This technique is not always feasible in large-scale programs because it is time consuming, and thus, costly. Interventions that require humans to score large quantities of responses may take inordinate amounts of time and deplete resources (Landauer et al., 2003; Williamson et al., 2012). Additionally, placing such potentially burdensome demands on human raters may increase the likelihood of fatigue and error (Ramineni & Williamson, 2013). In response to the practical limitations associated with manual hand scoring, automated scoring techniques may offer a possible solution.

## Automated Scoring Techniques

Automated scoring systems, in which responses are scored by machines, have been used to assess short-answer responses, essays, and spoken responses. For example, the Educational Testing Service (ETS) has utilized automated essay scoring (AES) for high-stakes assessments, such as the GMAT or GRE, for over 10 years (Williamson et al., 2012). Various AES systems exist, which all require large numbers of essay samples to base their scoring and feedback on. These systems tend to provide both holistic and specific feedback, although the exact content varies by system. Overall, AES systems have been found to be valid and reliable (Dikli, 2006).

Automated scoring techniques are not without controversy, however. In particular, AES has been criticized for oversimplifying the assessment of writing to focus on rote elements, such as word count or complexity of word choice, rather than less easily quantifiable aspects, such as thoughtfulness of response or writing to a specific audience (Condon, 2013; Perelman, 2014). The application of automated scoring techniques for short answer responses is less highly contested, especially when such techniques are not used to holistically rate the quality of writing with high-stakes implications. However, it has been suggested that automated short answer scoring is more arduous to create than AES, because AES tends to focus on grammar and mechanics while automated short answer scoring tends to focus on content (Brew & Leacock, 2013). Brew and Leacock (2013) further suggested that automated short answer scoring systems are underutilized because "it is currently impossible to buy an off-the-shelf short answer scoring engine that will work for all items" (p. 151).

Despite the aforementioned challenges, automated short answer systems are best suited to measure explicit concepts, facts, or skills (Brew & Leacock, 2013). Perhaps the most well-known automated system for short answer responses is known as "C-rater" which was developed by ETS (Leacock & Chodorow, 2003). The validity of C-rater has been evaluated by comparing automated scores with hand scores. Leacock and Chodorow (2003) found that automated scores matched scores assigned by human raters 84% of the time. Similar findings have been replicated elsewhere (Burstein, Chodorow, & Leacock, 2003), suggesting that automated short answer systems can offer a valid means of assessment. Despite the fact that the technology for creating and using similar automated systems has existed for at least a decade, few have been developed and disseminated across fields. Building off previous work, this study focused on the VAIL assessment, described in more detail below. The VAIL relies on short answer responses, rather than essays, and appears to be particularly conducive to adaptation into an automated system.

## Video Assessment of Instructional Learning (VAIL)

The VAIL is grounded in social learning theory, the notion that learning occurs largely through observation, as well as evidence suggesting that observational skills are valuable in developing expertise (Bandura, 1986; Jamil et al., 2015; Miller, 2011). The VAIL assesses observational skill by first asking teachers to watch a short video clip of an actual classroom and then identify and describe the effective teaching behaviors they observed. The process of "seeing" is an integral part of intentional teaching. One must be able to objectively identify and assess the effectiveness of specific practices in the classroom, and subsequently reflect on and modify personal teaching practices as necessary (Hamre, Downer, Jamil, & Pianta, 2012).

The content focus of the VAIL is teacher-child interactions, a topic particularly relevant to education. Teacher-child interactions have been consistently implicated as a means of promoting positive development in children (Burchinal et al., 2008; Thomason & LaParo, 2009; Yoshikawa et al., 2013). Subsequently, professional development has increasingly focused on training teachers to engage in positive interactions with children (Bierman, Nix, Greenberg, Blair, & Domitrovich, 2008; Domitrovich, Gest, Gill, Jones, & DeRousie, 2009; Pianta, Mashburn, Downer, Hamre, & Justice, 2008).

The VAIL has been validated (Jamil et al., 2015) and utilized in studies of both in-service (Hamre, Pianta, et al., 2012) and pre-service (Wiens, Hessberg, LoCasale-Crouch, & DeCoster, 2013) educators, offering a potentially valuable tool for teacher training and professional development. In a sample of pre-service teachers, demographic and programmatic characteristics did not consistently predict VAIL scores, suggesting that variation in VAIL scores were due largely to individual differences rather than group membership (Wiens et al., 2013). Furthermore, Jamil et al. (2015) found that the VAIL related to teachers' observed practices, suggesting that teachers who were more adept at identifying effective teacher-child interactions were also more likely to implement high quality teaching practices.

Furthermore, the National Center for Research on Early Childhood Education (NCRECE) Professional Development study, a randomized controlled evaluation of two forms of professional development, coursework and coaching, utilized the hand-scored VAIL as a measure of teachers' observational skills. At the end of the intervention, teachers enrolled in coursework were found to have significantly improved observational skills than teachers not enrolled in the course, and these improvements translated into meaningful changes in teachers' practice (Hamre, Pianta, et al., 2012; Downer et al., in press). Put differently, the VAIL was sensitive to intervention effects in the NCRECE course; this assessment tool seemed to detect important material teachers learned in the course that ultimately led to demonstrated improvements in practice. On the contrary, teachers' observational skills did not significantly change for teachers enrolled in the *coaching* intervention (Downer et al., in press). Although the reason for this is unclear, it may have been the result of the differences in content and delivery. For instance, it is possible that the VAIL may be more proximal to the content of the course, which focused on observing other teachers' practices, as opposed to the coaching intervention, which focused mostly on observing teachers' own personal practices.

In summary, although the VAIL measure is especially pertinent to current trends in education, the previously utilized hand scoring techniques may limit the scalability of the measure. As a result, the development of an automated means of scoring the VAIL was warranted. Automated scoring systems are indefatigable, systematic, and reliable, and offer a sustainable alternative to hand scoring. Nevertheless, it is necessary to ascertain that automated scoring systems are valid and useful, which were the aims of the present study.

## Current Study

The present study explored the extent to which the VAIL, an assessment of teachers' observational skills of teacher-child interactions, could be adapted into an automated scoring system. Specifically, the goals were to determine whether the automated VAIL scoring system related to the previously validated hand scoring system; and was sensitive to intervention effects

in the NCRECE professional development study. Building off of previous work utilizing hand-scores from the same study (Hamre, Pianta, et al., 2012; Downer et al., in press), we anticipated that the automated VAIL would be sensitive to intervention effects in the NCRECE course, but not the coaching intervention.

## Method

### Study Overview and Participants

This study utilized data from the NCRECE randomized, controlled evaluation of two forms of professional development designed to improve prekindergarten teachers' interactions with children. The NCRECE study was designed to evaluate scalable approaches to early childhood professional development (Pianta, Hamre, & Hadden, 2012). Teachers were placed randomly into treatment or control groups for the first phase, a 14 week (one semester) in-person course, and their group placements were then re-randomized for the second phase, the year-long MyTeachingPartner web-mediated coaching intervention.

The present study utilizes data on all teachers across all conditions who completed the post-intervention survey ($n$ = 175). Seventy-two (41.1%) teachers received the course in phase I and 88 (50.3%) received coaching in phase II. Most teachers (95.9%) were female, and 48.5% of all teachers were African American, 34.5% White, 10.5% Hispanic, 4.1% multi-racial, and 2.3% Asian. Roughly half (50.9%) of teachers taught in Head Start centers and 37.1% worked in public schools. On average, teachers were 41.54 years old ($SD$ = 10.41) with 10.93 years of experience teaching pre-kindergarten ($SD$ = 7.64). In terms of their degree attainment, 34.5% of teachers held a bachelor's degree, 33.3% had less than a bachelor's degree, and 32.2% held an advanced degree.

It is important to note that the current sample represents a fraction (43.5%) of the entire NCRECE sample ($n$ = 402). This reduced rate of completion is likely due to the fact that the post-intervention survey was optional. A series of t-tests were conducted to test whether those who completed the survey differed from those who chose not to complete the survey. Teachers did not significantly differ in terms of gender, age, and study condition; however, teachers who completed the survey were more likely to hold at least a bachelor's degree ($M$ = .65, $SD$ = .47) than those who did not complete the survey ($M$ = .58, $SD$ = .49; $t$ (378) = -1.5, $p$ <.01.) Teachers who completed the survey were also less likely to be Latino ($M$ = .10, $SD$ = .30) than those who did not complete the survey ($M$ = .19, $SD$ = .39; $t$ (378) = 2.31, $p$ <.001.)

The post intervention survey was sent via email and completed online. A portion of the survey included the Video Assessment of Interactions and Learning (VAIL) which is based on the Classroom Assessment Scoring System, or CLASS (Pianta, La Paro, & Hamre, 2008) a commonly used observational tool of teacher-child interactions. There are multiple forms of the VAIL focused on different elements of teaching; the present study focused on the VAIL designed to capture teachers' ability to detect aspects of Emotional Support, which includes creating a positive classroom climate, being sensitive to students' needs, and having regard for students' perspectives. To complete the VAIL, teachers viewed a two-and-a-half-minute video clip that depicted a teacher engaging a student in a conversation about her weekend. Teachers were instructed to watch the video as many times as they wished, but were encouraged not to

spend more than 10 minutes on the video. Then teachers were asked to name up to five strategies the teacher used to support the student's social and emotional development. Finally, teachers were asked to, list a specific, behavioral example of each strategy from the clip.

**Measures**

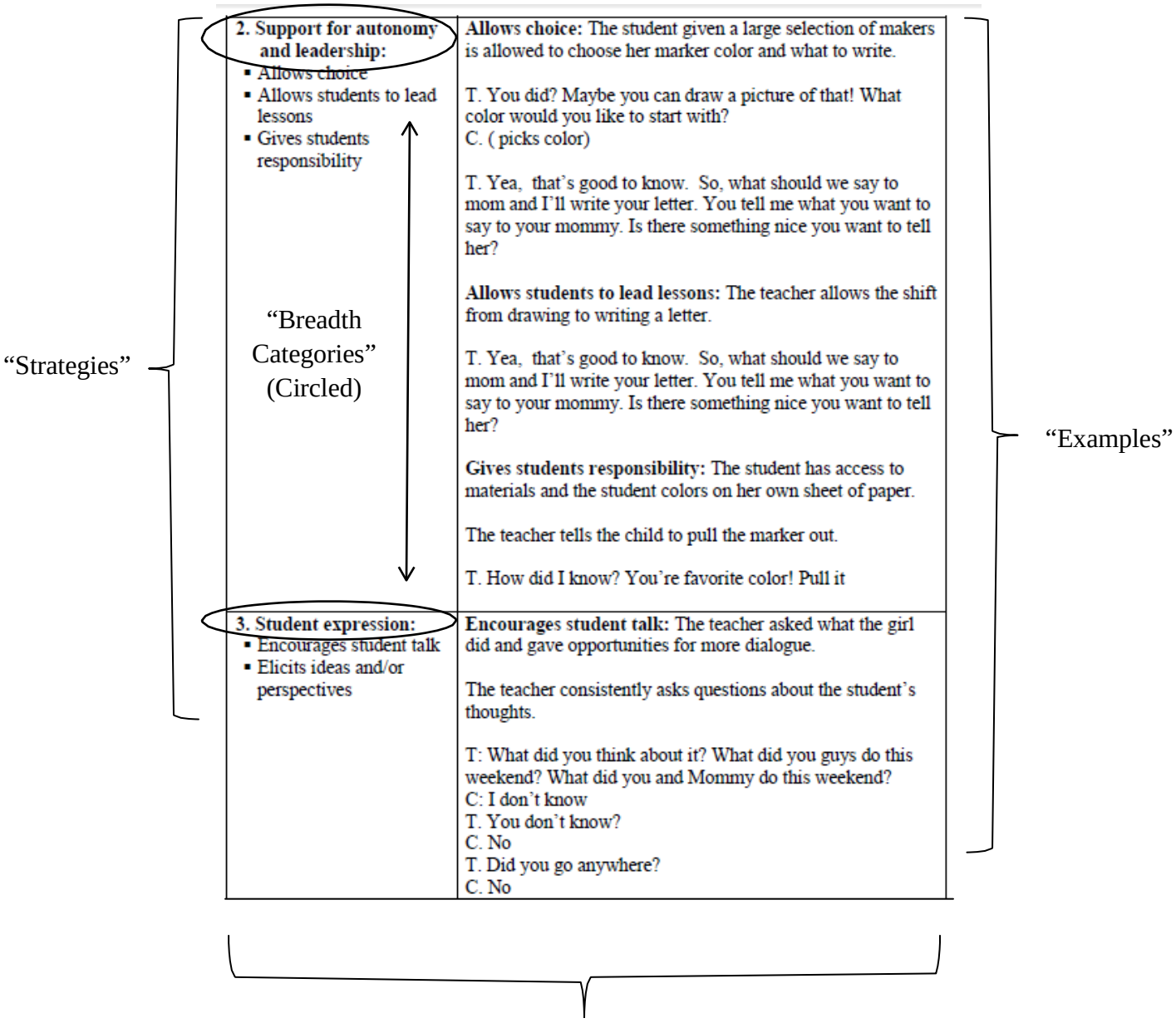
**Hand Scoring System.** All VAIL responses were hand scored based on a previously established standardized rubric that aligns with the CLASS (Pianta, La Paro, et al., 2008). For each response, trained coders assessed four elements: (1) *strategy*, if the learner identified a behavioral indicator consistent with the CLASS; (2) *example*, if the learner provides a specific behavioral description from the video of the teacher demonstrating a strategy; (3) *breadth*, the specific CLASS indicator that the strategy is most consistent with; and (4) *match*, whether the strategy and example pair is representative of the same CLASS indicator.

In keeping with the behavioral indicators for the Emotional Support domain of CLASS (Pianta, La Paro, et al., 2008), VAIL responses could fall into twelve possible *breadth* categories, including: (1) *Relationships*: being in close physical proximity, engaging in shared activity, matching the child's affect, and engaging in social conversation with the student; (2) *Positive Affect*, smiling, laughing, showing enthusiasm; (3) *Positive Communication:* demonstrating verbal affection (4) *Respect*: maintaining eye contact, having a warm and calm voice, cooperating and sharing; (5) *Awareness:* anticipating problems, noticing difficulties; (6) *Responsiveness*: acknowledging emotions, providing comfort and individualized support; (7) *Addressing Problems:* helping in a timely and effective manner and resolving problems; (8) *Student Comfort*: the student seeks support, freely participates, and takes risks; (9) *Flexibility and Student Focus:* showing flexibility, incorporating the student's ideas, following the student's lead; (10) *Support for Autonomy and Leadership:* allowing choice, allowing students to lead lessons, giving students responsibility; (11) *Student Expression*: encouraging student talk and eliciting ideas; and (12) *Restriction of Movement:* allowing movement, not being rigid.

Collectively, this information was then used to calculate, per teacher, the total number of: correct *strategies*, correct *examples*, unique *breadth* scores, and strategy-example *matches*. Because teachers could enter up to five possible responses, the range for each of the four aforementioned categories was 0-5. The four categories rendered a Cronbach's alpha coefficient of .79. An excerpt from the coding manual is shown in Figure 1 along with further description of how the manual was used to assess the four components (*strategy, example, breadth,* and *match*) that were described above.

**Automated Scoring System.** The VAIL hand coding manual was used to build a "dictionary" for the automated system. The dictionary was organized in a format that could be utilized in automated scoring software, specifically the Linguistic Inquiry Word Count (LIWC) software (Pennebaker, Booth, & Francis, 2007). In an attempt to replicate the hand scoring system, key words and phrases for the automated dictionary were pulled directly from the hand coding manual. For instance, according to the VAIL manual the teacher demonstrates regard for the student's perspective by allowing choice, so the term "choic*" was included in the automated dictionary. As part of the LIWC software, any letters in the word that appeared *after* the asterisk were disregarded (therefore "choice" and "choices" would both be considered correct.) Additionally, notes were collected from trained VAIL coders regarding language used in teacher

responses that varied from the coding manual, but was still considered correct (i.e., synonyms). Consistent with the previous example, words such as “choose”, “choosing” and “option” were included as synonyms of “choice”.




| | |
|---|---|
| **2. Support for autonomy and leadership:**<br>▪ Allows choice<br>▪ Allows students to lead lessons<br>▪ Gives students responsibility | **Allows choice:** The student given a large selection of makers is allowed to choose her marker color and what to write.<br><br>T. You did? Maybe you can draw a picture of that! What color would you like to start with?<br>C. ( picks color)<br><br>T. Yea, that's good to know. So, what should we say to mom and I'll write your letter. You tell me what you want to say to your mommy. Is there something nice you want to tell her?<br><br>**Allows students to lead lessons:** The teacher allows the shift from drawing to writing a letter.<br><br>T. Yea, that's good to know. So, what should we say to mom and I'll write your letter. You tell me what you want to say to your mommy. Is there something nice you want to tell her?<br><br>**Gives students responsibility:** The student has access to materials and the student colors on her own sheet of paper.<br><br>The teacher tells the child to pull the marker out.<br><br>T. How did I know? You're favorite color! Pull it |
| **3. Student expression:**<br>▪ Encourages student talk<br>▪ Elicits ideas and/or perspectives | **Encourages student talk:** The teacher asked what the girl did and gave opportunities for more dialogue.<br><br>The teacher consistently asks questions about the student's thoughts.<br><br>T: What did you think about it? What did you guys do this weekend? What did you and Mommy do this weekend?<br>C: I don't know<br>T. You don't know?<br>C. No<br>T. Did you go anywhere?<br>C. No |

Figure 1. Excerpt from the VAIL Coding Manual

The LIWC software program allows the user to upload a dictionary organized broadly by a designated number of categories with various words comprising each category. As a result, our dictionary contained a total of 24 categories that represented the 12 possible breadth categories for strategies and, similarly, the 12 possible categories for examples. Thus, further building on

the example outlined in Figure 1, in the automated dictionary, the *strategy* category for the "Support for Leadership and Autonomy" would be comprised of words such as "choice" and "responsibility". The corresponding *example* category would be comprised of phrases, such as "choose the marker" or "say to mommy." As previously mentioned, each teacher had up to five VAIL responses; to maintain consistency with the hand scoring technique, each response (one response per text file) was independently run through the software. The process of actually running all responses through the software was done simultaneously (all text files were run at the same time), allowing all responses to be scored in a matter of minutes.

LIWC provides output indicating the proportion of words that fall into certain categories, which, in this case, was used to assess whether or not the words used fell into any of the 24 previously mentioned categories. This information was then used to create: (1) *strategy* scores, represented as "correct" if a number other than zero appeared in any of the 12 strategy categories (suggesting that a word in the prescribed dictionary was used) (2) *example* scores, similarly, were "correct" if a number other than zero appeared in any of the 12 example categories (3) *breadth* scores, the specific category for which the strategy was located (4) *match* scores, if the strategy category matched the example category. Similar to the hand scoring technique, results ranged from 0-5. The four components rendered a Cronbach's alpha coefficient of .81. In addition, a mean score was also created by averaging across the four components in order to maintain consistency with previous work (Jamil et al., 2015).

## Results

First, descriptive statistics (Table 1) were examined to compare the various components (i.e., strategy, example, breadth, and match scores) of the automated scoring system and the hand scoring system. The means, ranges, and overall distributions proved to be very similar. The automated scoring system did not consistently yield higher or lower results than the hand scoring system.

Table 1
*Descriptive Statistics*

| | Mean | Std. Dev. | Range |
|---|---|---|---|
| Automated: Strategies | .86 | 1.26 | 0-5 |
| Hand-scored: Strategies | 1.05 | 1.35 | 0-5 |
| Automated: Examples | 3.60 | 1.52 | 0-5 |
| Hand-scored: Examples | 3.51 | 1.50 | 0-5 |
| Automated: Breadth | .85 | 1.11 | 0-4 |
| Hand-scored Breadth | .87 | 1.04 | 0-4 |
| Hand-scored: Match | .65 | 1.05 | 0-5 |
| Hand Code: Match | .78 | 1.18 | 0-5 |

Next, the four components of the automated system were correlated with the four components of the hand scoring system (Table 2). The bivariate correlations were consistently

high across the four components, ranging from .72 to .87, suggesting that the automated scoring system was consistently replicating the hand scoring. High correlations were also observed among the strategy, breadth, and match components for both the automated and hand-scoring systems, which is not surprising given that breadth and match scores are only assigned when a strategy is "correct", rendering high correlations among these elements. To further unpack these associations, Table 3 provides the percentage of cases in agreement and disagreement (by one, two and three or more points) among the automated and hand-scored systems. The automated and hand scoring systems agreed in roughly two thirds of all cases.

Table 2

*Correlations among Automated and Hand Scoring System*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Automated: Strategies | 1 | | | | | | | |
| 2. Hand-scored: Strategies | .82* | 1 | | | | | | |
| 3. Automated: Examples | .24* | .22* | 1 | | | | | |
| 4. Hand-scored: Examples | .15 | .15* | .87* | 1 | | | | |
| 5. Automated: Breadth | .92* | .74* | .24* | .15* | 1 | | | |
| 6. Hand-scored: Breadth | .75* | .94* | .21* | .13 | .73* | 1 | | |
| 7. Automated: Match | .92* | .79* | .28* | .19* | .85* | .72* | 1 | |
| 8. Hand-scored: Match | .70* | .89* | .28* | .31* | .64* | .81* | .72* | 1 |

Note: * = $p < .05$

Table 3

*Percentage of Cases in Agreement and Disagreement among Automated and Hand Scoring Systems*

| | Exact Match | Disagree by 1 | Disagree by 2 | Disagree by 3+ |
|---|---|---|---|---|
| Strategy | 65.7% | 25.7% | 8.0% | 0.6% |
| Example | 62.9% | 30.9% | 5.1% | 1.1% |
| Breadth | 66.3% | 24.6% | 9.1% | 0.0% |
| Match | 65.1% | 26.3% | 6.9% | 1.7% |

Last, we explored the extent to which the automated system could be used to detect intervention effects. As previously mentioned, the first intervention phase was a course, and the second phase was a coaching intervention. As a result, we tested the extent to which the automated scoring system could be used to detect intervention effects by comparing the means

for the treatment and control groups separately by phase (Table 4). For the course phase, three of the four components, as well as the mean score of the automated VAIL were found to be significantly different by intervention group. In particular, there were significant differences in scores for the treatment and control groups for strategy, breadth, match, and overall scores, such that the treatment group was significantly more likely to receive credit for these VAIL components than the control group. Associated effect sizes were moderately large, ranging from .54 to .60 (Table 4). There was no significant difference between groups for example scores. Consistent with findings from prior research using the hand-scoring method, there were no significant mean differences for any of the VAIL scores for the coaching phase. As shown in Table 4, the automated and hand-scored results yielded relatively similar effect sizes. In sum, these results suggest that the automated VAIL system replicated findings from the hand-scoring method.

Table 4

*Mean Differences and Effect Sizes by Treatment Condition& Intervention Phase*

| | | | Automated | | | Hand |
|---|---|---|---|---|---|---|
| | *t* | *p* | *M (SD)* Treatment | *M (SD)* Control | Cohen's *d* | Cohen's *d* |
| Phase I | *df* = 145 | | *n* = 72 | *n* =75 | | |
| Strategy | 3.50 | <.001 | 1.29 (1.61) | .56 (.81) | .57 | .56 |
| Example | .49 | .89 | 3.65 (1.52) | 3.53 (1.46) | .08 | .05 |
| Breadth | 3.57 | <.001 | 1.21 (1.30) | .57 (.81) | .59 | .55 |
| Match | 3.69 | <.001 | 1.02 (1.34) | .37 (.67) | .61 | .49 |
| Total | 3.29 | <.001 | 1.79 (1.19) | 1.26 (.72) | .54 | .51 |
| Phase II | *df* = 173 | | *n* = 88 | *n = 87* | | |
| Strategy | .36 | .55 | .93 (1.17) | .79 (1.35) | .11 | .07 |
| Example | .66 | .42 | 3.55 (1.57) | 3.66 (1.49) | -.07 | -.07 |
| Breadth | .14 | .71 | .93 (1.09) | .76 (1.12) | .18 | .14 |
| Match | 1.16 | .28 | .66 (.87) | .63 (1.20) | .02 | .04 |
| Total | .38 | .97 | 1.52 (.96) | 1.45 (1.04) | .07 | .06 |

## Discussion

Automated scoring systems offer a systematic and sustainable alternative to hand scoring techniques, which may be particularly valuable in large-scale interventions (Shermis & Burstein, 2013). The purpose of this study was to explore the extent to which an automated VAIL scoring system was able to replicate scores produced by the previously validated hand scoring system. Findings show that the automated VAIL scores were strongly related to the hand coded VAIL scores, suggesting that the automated VAIL consistently replicated previous scores. Similarly,

the automated VAIL system detected similar intervention effects as the hand-scoring method. Taken together, these findings suggest that the automated VAIL appears to be a valid and reliable means of measurement.

It is worth noting that the strength of the associations between automated and hand-scored systems varied by what was assessed. More specifically, strategy and example scores had the strongest associations while match and breadth scores had the weakest. This is not surprising given that the automated dictionary was specifically built around the language of strategies and examples. In the VAIL, breadth and match are only scored if the strategy is found to be "correct", as neither breadth nor match are relevant if the strategy is incorrect. This has implications in our study because exact matches of *breadth* and *match* among automated and hand coding systems required an exact match on *strategy* as well. This likely accounts for the slightly higher discrepancy among the two systems for breath and match scores, although the correlations are still high.

The automated system replicated roughly two thirds of hand scores, which is slightly lower than previous work (Leacock & Chodorow, 2003). Of course, exact replication of the hand scored system is irrelevant in building an automated system given that the hand scores contain some degree of human error. Instead, we sought to develop a better system for measurement that would address issues in scalability.

Utilizing the automated VAIL or similar measures may be particularly advantageous in large-scale interventions, such as online courses, which are pervasive in various fields (U.S. Department of Education, National Center for Education Statistics, 2016). Automated measures allow intervention designers and instructors to incorporate open-ended questions into learner assessments, rather than simply relying on the more restrictive, close-ended alternative. Previously, open-ended questions posed logistical challenges for scoring in large-scale programs; however, automated systems offer a feasible alternative to manual hand scoring (Shermis & Burstein, 2013). Aside from open-ended questions being a better metric of student learning, this solution is also cost-effective and efficient (Williamson et al., 2012). The time spent designing and using an automated system is front-loaded, meaning more time is spent in the construction of the system than in the actual process of running the system to obtain scores (Brew & Leacock, 2013). By comparison, the hand scoring systems require more time to be spent maintaining highly reliable coders. Although human coders are likely to become more efficient over time, the possibility of making a mistake is ever-present (Ramineni & Williamson, 2013).

Although the VAIL is most applicable to teacher education, the findings from this study may be relevant to other fields. For this reason, it is useful to consider why the VAIL was particularly conducive to adaptation and how this information can be used in other studies. In this study, correct responses were finite and language-specific. More specifically, there were a fixed number of correct strategies that teachers could identify, and thus, the building of the automated dictionary encompassed words taken directly from the pre-constructed manual, as well as a limited number of associated synonyms. This was due in large part to the fact that VAIL responses were prompted by the viewing of a short video clip followed several short-answer questions. In other words, respondents were not asked to generate examples of Positive Climate from their potentially limitless repertoires of personal experience; instead, they were

asked to identify concrete, specific examples from the video. Furthermore, there is less controversy surrounding the use of automated systems for scoring short answer responses, as compared to essays. AES has been criticized for oversimplifying the assessment to abstract concepts that may be difficult to quantify (Condon, 2013; Perelman, 2014).

Course designers in various fields could use relevant video, short writing excerpts, or images to elicit explicit concepts or facts conducive to automated scoring. Consistent with the use of videos to demonstrate how to *do* (or not do) something, videos could also be used to assess a learners' abilities to effectively *see* a skill or concept in action (Hamre, Downer, et al., 2012). For instance, kinesiology students could watch a video of a person exercising and analyze his gait or specific muscles that are being activated. Assuming that the video could be used to generate a finite number of correct responses, this information lends itself well to automated scoring, which would be particularly useful if a large number of students were enrolled in the course. In other fields, such as English or history, it may be most relevant to utilize writing excerpts or images to test learners' understanding of specific content. Ultimately, as online coursework continues to be scaled up, across fields, it is necessary to think creatively about how to effectively assess learning. Automated scoring systems offer one possible technique, especially for short-answer assessments that have a limited number of correct answers.

**Limitations**

In the present study, VAIL scores for the CLASS domain of interest were only collected at the conclusion of the intervention study. As a result, it was not possible to examine changes in VAIL scores from the beginning to the end of the intervention. In addition, the timing and delivery of the VAIL at the end of the post-intervention survey may have contributed to a lowered response rate. Those who self-selected to complete this survey proved to be different than those who chose not to complete the survey in terms of ethnicity and educational attainment.

**Future Directions**

Future work should continue to develop and test automated systems across fields. Given the availability of various automated software programs and the ecological validity of using automated systems, it is important for researchers to share their methods in developing and testing automated systems. The process of automating assessment scoring should be driven by the characteristics of the assessment questions and answers. Therefore, sharing techniques for automating various types of assessments, and establishing validity, can advance the use of automated scoring systems across fields.

In terms of the automated VAIL specifically, it would be beneficial to conduct similar inquiries using different samples of teachers enrolled in different interventions. Future studies should continue to utilize different samples to test the validity of the automated VAIL to verify that the current results consistently replicate. In light of recent findings that suggest automated scoring techniques differentially benefit native English speakers (Reilly et al., 2016), it would be advantageous to test the automated VAIL, and similar measures, in linguistically diverse samples to ensure validity. As previously mentioned the present study considered in-service teachers who were enrolled in coursework and/or coaching. Given the utility of the automated VAIL in large-scale interventions, future inquiries should also examine the extent to which the automated VAIL can be used to detect intervention effects using other intervention designs, such as online

coursework. Additionally, based on previous work suggesting the VAIL is valuable in assessing teacher skill in both in-service and pre-service teachers (Wiens et al., 2013), it would also be informative to utilize samples of pre-service, as well as in-service teachers.

## Conclusion & Implications

The design and delivery of scalable interventions and programs is of central importance to many fields utilizing online coursework. In teacher education, a variety of professional development programs for teachers have focused on training teachers to engage in positive interactions with children (Bierman et al., 2008; Domitrovich et al., 2009; Pianta, Mashburn, et al., 2008). However, such large-scale interventions pose challenges to assessment, especially the assessment of more complex open-ended responses. As a result, the automated VAIL appears to offer a valid and practically useful means of assessing teachers' skills in observing teacher-child interactions with implications for pre-service training, higher education, and professional development.

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B090002 and R305A060021 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the funding agencies.

## References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology, 20,* 821-843. doi: 10.1017/S0954579408000394

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university*. New York: McGraw-Hill International.

Boston, C. (2002). The concept of formative assessment. ERIC Digest. Retrieved from ERIC database. (ED470206).

Brew, C. & Leacock, C. (2013). Automated short answer scoring: Principles and prospects. In M. D. Shermis & J. Burstein (Eds.) *Handbook of automated essay evaluation: Current applications and new directions* (pp. 136-152). New York: Routledge.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*(3), 140-153. doi: 10.1080/10888690802199418

Burstein, J., Chodorow, M. & Leacock, C. (2003). Criterion online essay evaluation: An application for automated evaluation of student essays. Retrieved from American Association for Artificial Intelligence: http://www.aaai.org/Papers/IAAI/2003/IAAI03-001.pdf

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100-108. doi: 10.1016/j.asw.2012.11.001

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1).

Domitrovich, C. E., Gest, S. D., Gill, S., Jones, D., & DeRousie, R. S. (2009). Individual factors associated with professional development training outcomes of the Head Start REDI Program. *Early Education & Development, 20*, 402-430. doi: 10.1080/10409280802680854

Downer, J. T., Pianta, R. C., Burchinal, M., Field, S., Hamre. B. K. …Scott-Little, C. (in press). Coaching and coursework focused on teacher-child interactions during language/literacy instruction: Effects on teacher beliefs, knowledge, skills, and practice.

Foddy, W. (1993). Constructing questions for interviews and questionnaires: Theory and practice in social research. Cambridge: Cambridge University Press.

Franks, R. P. & Schroder, J. (2013). Implementation science: What do we know and where do we go from here? In T. Halle, A. Metz, & I. Martinez-Beck (Eds.) Applying implementation science in early childhood programs and systems (pp 5-19). Brookes Publishing: Baltimore.

Gill, W. E. (2011). The Ready to Teach program. A federal initiative in support of online courses for teachers. Retrieved from http://files.eric.ed.gov/fulltext/ED530966.pdf

Hamre, B. K., Downer, J. T., Jamil, F. M., & Pianta, R. C. (2012). Enhancing teachers' intentional use of effective interactions with children. In R. C. Pianta (Ed.) *Handbook of early childhood education* (pp 507-532). New York: The Guilford Press.

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., . . . Scott-Little, C. (2012). A Course on Effective Teacher-Child Interactions: Effects on Teacher Beliefs, Knowledge, and Observed Practice. *American Educational Research Journal*, 88-123. doi: 10.3102/0002831211434596

Jamil, F. M., Sabol, T. J., Hamre, B. K., & Pianta, R. C. (2015). Assessing teachers' skills in detecting and identifying effective interactions in the classroom: Theory and measurement. *The Elementary School Journal, 115*(3), 407-432. doi: 10.1086/680353

Landauer, T. K., Laham, D. & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10*(3), 295-308. doi: 10.1080/0969594032000148154

Leacock, C. & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*, 389-405.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. Retrieved from U.S. Department of Education website: https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf

Miller, K. (2011). Situation awareness in teaching: What educators can learn from video-based research in other fields. In M. Sherin, V. Jacobs, & R. Phillip (Eds.) *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 51-65). New York: Routledge.

Palloff, R. M. & Pratt, K. (2008). *Assessing the online learner: Resources and strategies for faculty.* San Francisco: Jossey-Bass.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC2007 Operator's Manual*. Retrieved from http://www.LIWC.net.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104-111. doi: 10.1016/j.asw.2014.05.001

Pianta, R. C., Hamre, B. K., & Hadden, D. S. (2012). Scaling up effective professional development. In C. Howes, B. Hamre, & R. Pianta (Eds.) *Effective early childhood professional development: Improving teacher practice and child outcomes* (pp.191-212). Baltimore: Brookes Publishing Company.

Pianta, R., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System.* Baltimore: Brookes Publishing Company.

Pianta, R., Mashburn, A., Downer, J. Hamre, B., & Justice, L. (2008). Effects of Web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23*(4), 431-451. doi: 10.1016/j.ecresq.2008.02.001.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practice. *Assessing Writing, 18*, 25-39. doi: 10.1016/j.asw.2012.10.004.

Reilly, E. D., Williams, K. M., Stafford, R. E., Corliss, S. B., Walkow, J. C., & Kidwell, D. K. (2016). Global times call for global measures: Investigating automated essay scoring in linguistically-diverse MOOCs. *Online Learning, 20*(2), 97-109.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions.* New York: Routledge.

Thomason, A.C. & La Paro, K.M. (2009). Measuring the quality of teacher-child interactions in toddler child care. *Early Education & Development, 20*(2), 285-304. doi: 10.1080/10409280902773351.

U.S. Department of Education, National Center for Education Statistics. (2016). *Digest of Education Statistics, 2014* (NCES 2016-006), Table 311.15.

Vale, K & Littlejohn, A. (2014). Massive open online courses: A traditional or transformative approach to learning? In A. Littlejohn & C. Pegler (Eds.) *Reusing open resources: Learning in open networks for work, life and education* (pp. 138-153). New York: Routledge.

Wiens, P., Hessberg, K., LoCasale-Crouch, J., & DeCoster, J. (2013). Using a standardized video-based assessment in a university teacher education program to examine pre-service teachers knowledge related to effective teaching. *Teaching and Teacher Education, 33*, 24-33. doi: 10.1016/j.tate.2013.01.010.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13. doi: 10.1111/j.1745-3992.2011.00223.x

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., & Zaslow, M. J. (2013). Investing in our future: The evidence base on preschool education. Retrieved from Society for Research in Child Development website: http://www.srcd.org/sites/default/files/documents/washington/mb_2013_10_16_investing_in_children.pdf