

DEVELOPMENT OF THE LARGE SCALE DIAGNOSTIC ASSESSMENTS OF COLLEGE
SKILLS

Fusun Sahin¹, Jason Bryer², Heidi Andrade³, Angela Lui³, David Franklin³, Diana
Akhmedjanova³

¹ American Institutes for Research

²Excelsior College

³University at Albany, State University of New York

Acknowledgement: The work was done while the first author was a doctoral student at the University at Albany, State University of New York.

Most institutions of higher education identify and address incoming students' preparedness by administering placement examinations in fundamental content areas (i.e., reading, writing, and mathematics). Placement exams can be helpful for identifying students who are at-risk of struggling in the academic areas tested, and for placing students in remedial courses. However, placement exams typically do not offer students individualized feedback on their academic strengths and weaknesses, nor do they link students to resources beyond the remedial courses. Furthermore, this approach to assessing students' preparedness for college work is associated with increased costs and time to degree, which lead to attrition (Bailey & Cho, 2010; Belfield & Crosta, 2012; Scott-Clayton, 2012). In a recent meta-analytical study, Valentine, Konstantopoulos, and Goldrick-Rab (2017) found that university students who took remedial reading and writing courses due to low scores on a placement exam earned fewer college credits than those who were not in the remedial programs.

Finally, typical placement exams do not measure important factors of student success, such as self-regulated learning (SRL), metacognition, motivation, self-efficacy, mindset, and grit, all of which have been found to be significant predictors of academic achievement (e.g., Donker, de Boer, Kostons, Dignath van Ewijk, & van der Werf, 2014; Duckworth, Peterson, Matthews, & Kelly, 2007; Lai, 2011; Zimmerman, 2011). These limitations of traditional placement examinations inspired the development of the Diagnostic Assessment and Achievement of College Skills (DAACS), a free, online, multidimensional, diagnostic assessment system of key skills, as well as individualized feedback and links to useful resources.

Purpose

The DAACS is an open-source, technology-based platform that provides no-stakes diagnostic assessments of academic and non-academic skills, as well as instant, actionable feedback. The aim of DAACS is to combat high dropout rates in higher education by measuring

incoming college students' skills in reading, writing, mathematics, and SRL, providing feedback on strengths and weaknesses, and sharing links to relevant learning resources. Moreover, academic advisors, mentors, and coaches are trained so that their advice to students in credit-bearing courses can draw on the feedback provided by DAACS.

The DAACS is currently being piloted at an online college. The purpose of this paper is to share preliminary results regarding the validity and reliability of the inferences drawn from each of the four DAACS assessments: math, reading, writing, and a self-report survey on SRL.

Theoretical Framework

A vital step in developing an instrument is to establish the validity and reliability of the inferences drawn from the instrument. Test validity is defined as “the degree to which evidence and theory support the interpretation of the scores for proposed uses” (AERA, APA, & NCME, 2014, p. 11). According to the unitary concept of validity (AERA, APA, & NCME, 2014; Messick, 1995), there are six aspects of construct validity evidence that indicate how well an instrument is measuring the construct it is intended to measure. Given that we are still in the pilot phase of the project, we have preliminary evidence of three aspects – test content, internal structure, and relations to other variables. We also have preliminary evidence on the reliability of our instruments.

According to Messick (1995), the first aspect of construct validity evidence is test content, indicating the relevance and representativeness of the construct to be assessed. In this case, the content of the tests of self-regulated learning, reading, writing, and math should be aligned with content suggested by theory and expert opinion, and relevant to adult learners in an online context.

The second aspect of construct validity evidence is internal structure, which indicates how well the relationships between assessment items and components correspond with the theory under which the assessments were developed. This means the factor structure of the SRL survey scales should represent metacognitive, behavioral, and motivational aspects of SRL (Zimmerman & Schunk, 2011), and correlations should suggest that these domains are distinct yet related to each other. Similarly, each of the criteria on the rubric used to score the writing assessment should represent distinct domains but should also be correlated with each other. Since the reading and mathematics assessments are designed to be unidimensional, we expect them to have high overall internal consistency.

The third aspect of construct validity evidence is the relations of scores on the instrument to other variables, which indicate the extent to which the patterns of association between and among scores on the test under study and other variables are consistent with theoretical expectations (Messick, 1995). Relationships between measures of the related constructs are expected to be high, representing convergent validity, while relationships with measures of distinct constructs are expected to be low, representing divergent validity. The subscales of the SRL survey (metacognition, learning strategies, and motivation) are expected to be highly correlated, while correlations between the SRL survey and math, reading, and writing scores are expected to be moderate. Moreover, various studies reported that SRL, metacognition, motivation, and learning strategies significantly predicted academic achievement (e.g., Donker, de Boer, Kostons, Dignath van Ewijk, & van der Werf, 2014; Lai, 2011; Zimmerman, 2011). Therefore, responses to the SRL survey are expected to correlate with various achievement indicators such as course grades, retention, and credits earned, which can serve as additional evidence of construct validity.

The reliability of a measure refers to the consistency or precision of scores (AERA, APA, & NCME, 2014). The most relevant type of reliability for the math, reading, and SRL assessments is internal consistency, which refers to the correlations between items within each assessment and domain. For the writing assessment, interrater reliability (i.e., consistency across raters) is most relevant since the assessment involves human and automated scoring. Interrater reliability between humans must first be established, followed by interrater reliability between human raters and machine scoring.

This paper addresses the three sources of construct validity evidence and reliability mentioned above for the four DAACS assessments: SRL Survey, writing, reading, and math. Specifically, the purposes of this paper are to report the validity and reliability evidence with regard to:

- content, which is based on reviews from content experts and advisors for the DAACS SRL survey, reviews from content experts and students for the DAACS writing assessment, and reviews from content expert reviews for the DAACS reading and mathematics assessments.
- internal structure, which is based on exploratory and confirmatory factor analyses for the DAACS SRL survey, and correlations between domains and subdomains for the DAACS writing assessment.
- relations, which are examined by the correlations between the four DAACS assessments and student achievement, which is operationalized as the number of courses students passed.
- reliability, which is calculated by using Cronbach's alpha for the DAACS SRL survey, mathematics, and readings assessments and by using inter-rater reliability for the DAACS writing assessment.

Method

We have developed and piloted the four DAACS assessments: the SRL survey, an adaptive math test, an adaptive reading test, and a test of writing that asks students to write a brief essay in which they reflect on the results of their SRL survey. Students' proficiencies in each of these four areas are reported to them in terms of three levels: emerging, developing, and mastery.

Participants

The sample included 4614 incoming students in a private, nonprofit, fully-online university in the Western region of the U.S. who enrolled between April and September 2017. Demographics for a part of our sample are provided in Table 1.

Table 1
Sample Demographics (n=2499)

Demographics		<i>n (%)</i>
Age	< 20	15 (0.6%)
	20–29	920 (36.8%)
	30–39	1016 (40.7%)
	40–49	435 (17.4%)
	50–59	100 (4.0%)
	60 +	13 (0.5%)
Ethnicity	Asian	83 (3.3%)
	Black or African American	274 (11%)
	American Indian or Alaska Native	13 (0.5%)
	White	1696 (67.9%)
	Hispanic	290 (11.6%)
	Unknown	55 (2.2%)
	Bi- or Multi-racial	74 (3.0%)
	Native Hawaiian or Other Pacific Islander	13 (0.5%)
College Experience	Nonresident Alien	1 (0%)
	First-generation college attendee	1037 (41.5%)
Military Status	Active or Veteran	17 (0.7%)
Income	<16,000	195 (7.8%)
	\$16,000 - \$24,999	217 (4.9%)
	\$25,000 - \$34,999	361 (8.1%)
	\$35,000 - \$44,999	357 (8.0%)
	\$45,000 - \$64,999	464 (10.5%)
	\$65,000 or more	757 (17.1%)

Prefer not to answer	2 (0%)
Not reported	146 (5.8%)

Data Sources and Procedures

The mathematics, reading, and SRL assessments were constructed with selected-response items which drew on existing tests and surveys. The writing assessment was constructed with a constructed-response item where students were prompted to write an essay based on the results of their SRL survey responses. The writing assessment is designed to not only provide an accurate measure of students' writing abilities, but also to amplify the impact of the SRL survey by having students reflect on their results and commit to improving their SRL skills. Various steps were taken to construct each assessment and ensure the validity of responses to each assessment. Figure 1 depicts the steps taken to develop the four assessments and collect sources of validity and reliability evidence.

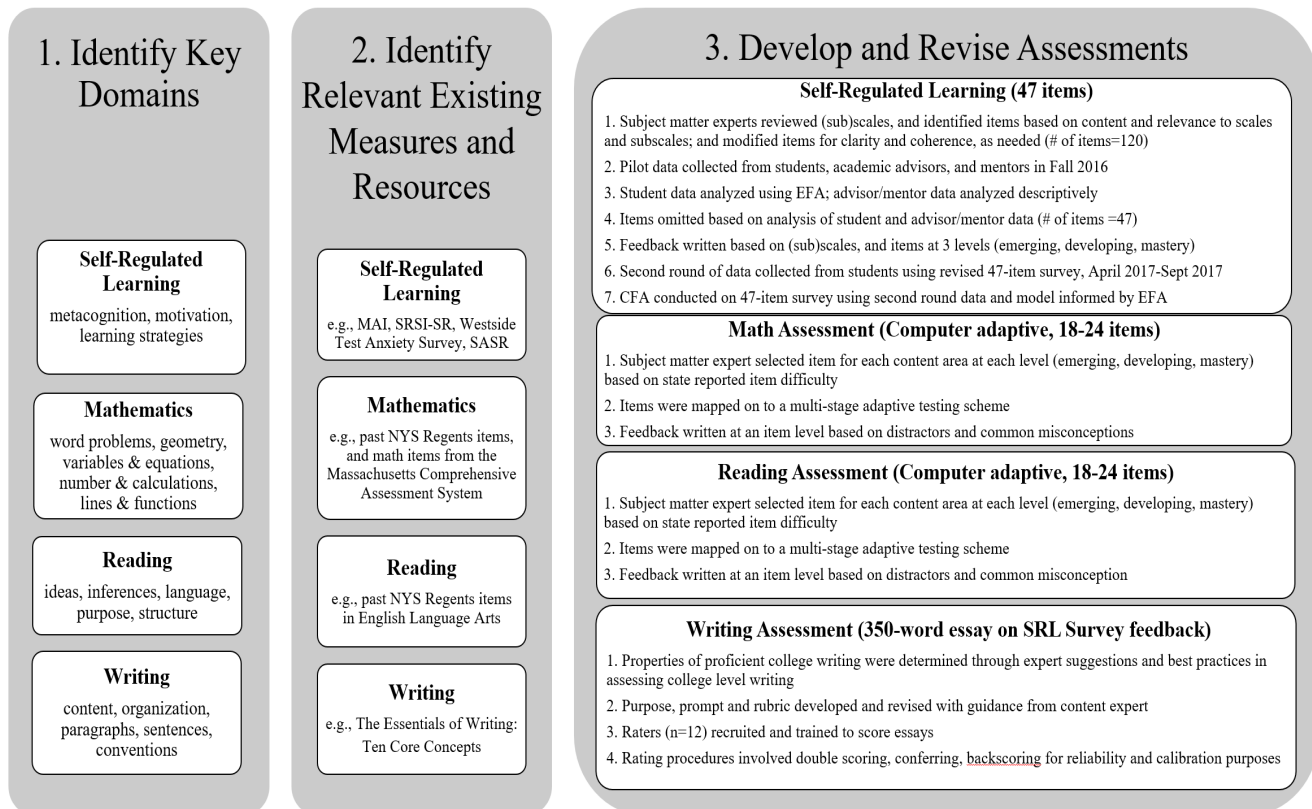


Figure 1. Steps taken to develop assessments for the DAACS

Self-regulated learning survey. SRL is “the processes whereby learners personally activate and sustain cognitions, affects, and behaviors that are systematically oriented toward the attainment of personal goals” (Zimmerman & Schunk, 2011, p. 1). In order to collect evidence that the SRL survey reflects this definition, content experts reviewed existing instruments and relevant constructs (e.g., metacognition, self-efficacy, motivation, and grit), and identified subscales and questions based on content, clarity, and usefulness of feedback based on these questions. One-hundred-and-twenty items were pilot-tested; 47 items were selected for the final survey.

Writing. The writing assessment was developed with four purposes in mind: (1) to give students targeted, actionable feedback about the critical elements of their writing; (2) to direct students toward relevant writing resources; (3) to assist students in reflecting on their DAACS

results and committing to a course of action related to their SRL; and (4) to enable academic advisors to review students' essays, which will supplement information obtained from the DAACS about students' strengths and weaknesses in terms of SRL. The properties of effective college-level writing were identified based on expert suggestions and best practices in assessing writing. Students were prompted to write an essay about their SRL results and feedback. A writing rubric was constructed with nine criteria based on established metrics of college writing and described writing proficiency in three levels: emerging, developing, and mastery. Experts evaluated the rubric and students were interviewed about the writing task and the criteria.

Twelve raters were trained by two content experts to score essays using the nine criteria rubric. Of the 2047 ratings thus far, 467 essays were scored by two raters, and the raters conferred when they disagreed. Twenty-five essays were also backscored by expert raters to check for drift. The essays and the scores for each essay that were agreed upon by two raters were used to train LightSide – an open-source automated essay-scoring algorithm. Scoring of the remaining essays ($n=3,207$) was performed by this algorithm.

Reading. Experts identified components of college-level reading. Items measuring these components were selected from a state-mandated high school English language arts exam. The questions were organized within reading passages and administered in a multi-stage computer-adaptive fashion. The items were placed into the testlets (i.e., each block of assessment at each stage of the multi-stage adaptive test) based on their reported difficulty parameters. Each student received between 18 and 24 items (see Figure 2 for the routing logic).

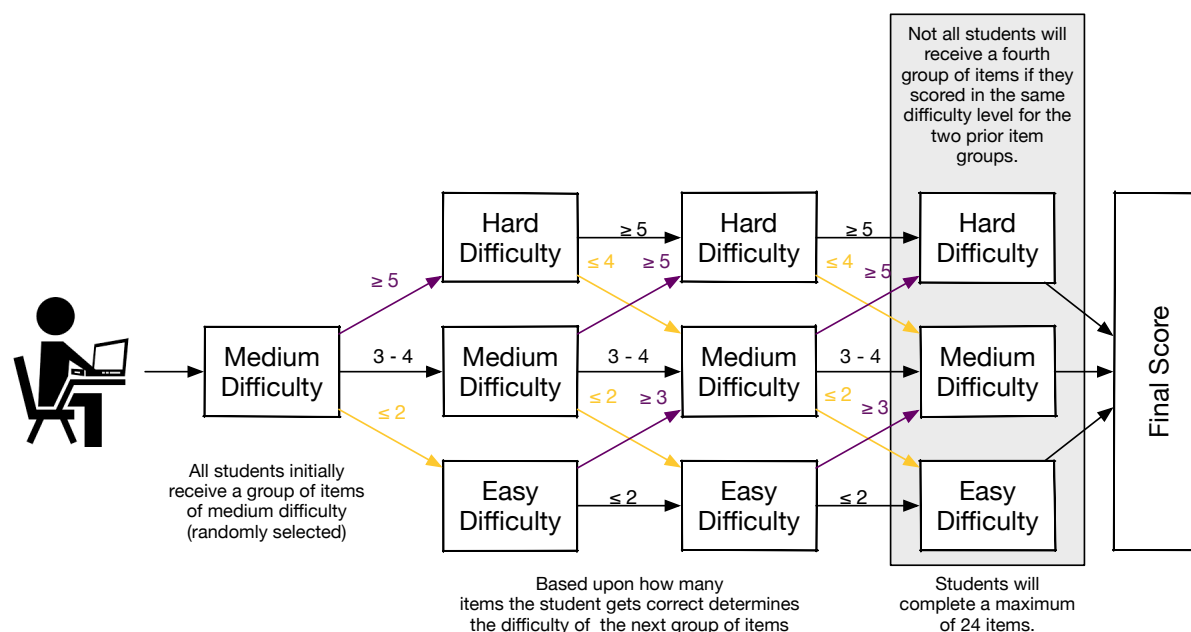


Figure 2. The multi-stage adaptive testing for the reading and math assessments.

Mathematics. Expert suggestions and existing instruments were used to define readiness for college-level mathematics. Questions were selected from state-mandated high school assessments from two states. The items were organized in testlets to prepare for the multi-stage adaptive administration. The items were placed into the testlets based on both the reported difficulty parameters of the items and their subcontent (e.g., algebra, statistics, geometry). Each student received between 18 and 24 items (see Figure 2 for the routing logic).

Results

Construct validity evidence reported in this paper is based on data collected from April 2017 to September 2017. Validity and reliability evidence for the four DAACS assessments are presented based on the sources mentioned in the theoretical framework: content, internal structure, relations between assessments, and reliability.

DAACS Self-Regulated Learning Survey

Test content. Because self-regulated learning (SRL) is a broad concept that includes students' cognition, affect, and behavior, the SRL survey was developed to encompass multiple dimensions. Four content experts identified 120 items from relevant existing measures (see Table 2), which were then sent to three other experts for review. After initial revisions, all seven experts came together to evaluate selected scales and subscales to ensure the content coverage of the SRL survey. Content experts focused on three aspects of each item: a) whether the item measured an important feature of self-regulated learning and the respective subconstruct on which the item loaded, b) whether there was another item that measured the same content, and c) whether the item lent itself to actionable feedback to students.

Table 2

Measures Used for the DAACS SRL Survey*

Scale	Factors (number of items)	Cronbach's Alpha
<i>Survey of Academic Self-Regulation (SASR) Self-Regulation</i>	1. Metacognition ($n=15$) 2. Self-Efficacy ($n=13$) 3. Self-Regulation ($n=10$) 4. Intrinsic Motivation ($n=7$) 5. Anxiety ($n=5$) 6. Extrinsic Motivation ($n=5$)	.86
<i>Self-Regulation Strategy Inventory – Self-Report (SRSI – SR)</i>	1. Managing Learning Environment and Behavior ($n=12$) 2. Seeking and Learning Information ($n=8$) 3. Maladaptive Regulatory Behavior ($n=8$)	.72 .88 .84
<i>Online Learning Value and Self-Efficacy Scale (OLVSES)</i>	1. Task Value ($n=6$) 2. Self-Efficacy ($n=5$)	.85 .87
<i>Metacognitive Awareness Inventory (MAI)</i>	1. Knowledge of Cognition ($n=17$): a. Declarative Knowledge ($n=8$) b. Procedural Knowledge ($n=4$) c. Conditional Knowledge ($n=5$) 2. Regulation of Cognition ($n=29$): a. Planning ($n=7$) b. Information Management Strategies ($n=8$) c. Monitoring ($n=7$) d. Debugging Strategies ($n=5$)	.88 .88

e. Evaluation ($n=2$)		Overall: .93
<i>Westside Test Anxiety Scale</i>	1. Impairment ($n=6$) 2. Worry and Dread ($n=4$)	N/A
<i>Mindset</i>	1. Fixed Mindset ($n=3$) 2. Growth Mindset ($n=3$)	.78
<i>Grit</i>	1. Consistency of Interest ($n=5$) 2. Perseverance of Effort ($n=5$)	.73-.83

*Permission was granted by authors of the original measures to use and adapt their items and scales.

After piloting the 120-items, experts – which included both SRL content experts and college advisors – selected items based on (a) whether the content of an item was essential to SRL; (b) lack of overlap between the content of other items; and (c) satisfactory item statistics and factor loadings. Forty-one items that met these criteria and had acceptable statistics remained in the survey. In addition, items related to mindset were included and remained intact in the survey due to their relevance, which made up 47 items for the SRL survey.

The version used for pilot testing consists of 47 Likert-scale type items with three anchor types: (1) 1 = *strongly disagree* to 5 = *strongly agree*, 2) 1 = *not like me at all* to 5 = *very much like me*, and (3) 1 = *almost never* to 5 = *almost always*. The DAACS SRL survey measures three SRL domains – metacognition, strategies, motivation. The breakdown of the number of items and descriptive statistics of students' responses to items in each domain are provided in Table 3.

Table 3

Number of Items and Descriptive Statistics for the Responses in SRL Survey Domains

Domains & Sub-domains	# of Items	Mean (SD)	SE	Median	Min	Max
Metacognition	13	34.7 (8.1)	0.12	35	0	52
Planning	3	8.2 (2.2)	0.03	8	0	12
Monitoring	6	16.3 (3.8)	0.06	17	0	24
Evaluation	4	10.1 (2.9)	0.04	10	0	16
Motivation		62.0 (8.6)	0.13	62	24	80
Anxiety	6	17.1 (4.6)	0.07	18	0	24
Self-efficacy	4	13.2 (2.1)	0.03	13	0	16
Mastery-goal orientation	4	13.3 (1.9)	0.03	13	0	16
Mindset	6	18.4 (3.8)	0.06	18	0	24
Strategies		42.3 (6.8)	0.10	43	12	56
For managing time	3	8.2 (2.0)	0.03	8	1	12
For help seeking	3	9.8 (2.0)	0.03	10	0	12
For managing environment	3	8.6 (2.2)	0.03	9	0	12
For understanding	5	15.7 (2.7)	0.04	16	0	20
TOTAL	47					

*The SRL survey uses a 5-point Likert-scale

Internal structure. In order to collect evidence of internal structure, exploratory factor analysis (EFA) was conducted for the 120-item version of the SRL survey, using the pilot data collected before April 2017 ($n=338$). Based on the factor loadings from EFA, a model was tested with the revised 41-item version using confirmatory factor analysis. This model had three second order constructs: motivation, metacognition, and strategies, and 11 first order latent variables: mastery orientation, mindset, self-efficacy, managing time, anxiety, evaluation, monitoring, managing environment, managing understanding, managing help-seeking, planning. The details are shown in Figure 3. The results of CFA indicated a good model fit, $\chi^2 (1020) = 21661.52$, $\chi^2/1020 = 13.83$; RMSEA = .053, 90%CI = [.052 to .054]; TLI = .86; CFI = .868; SRMR = .052.

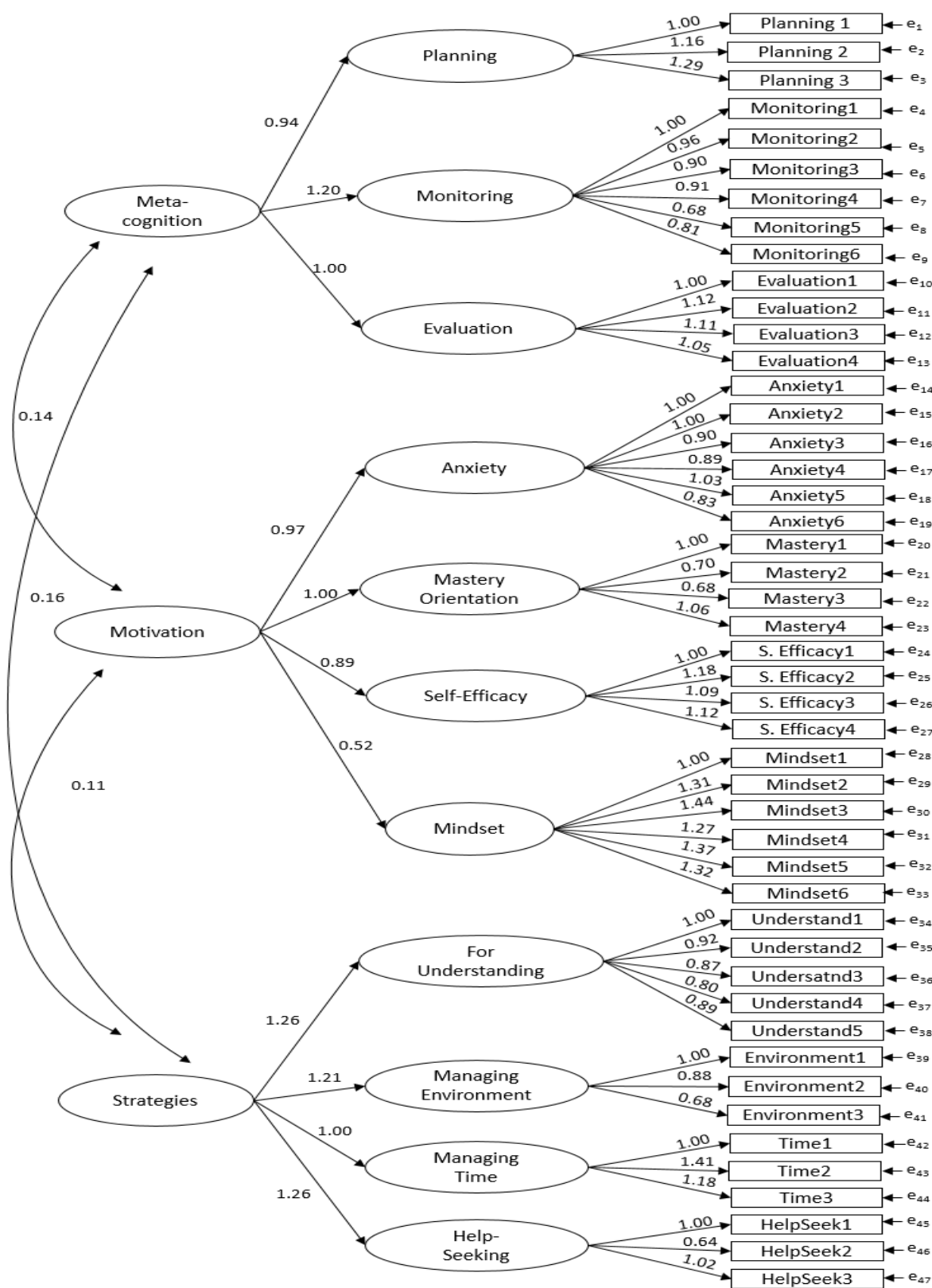


Figure 3. Graphic representation of DAACS SRL survey model (Item names are mapped to actual items in Appendix A).

Correlations between the subdomains, suggested by the EFA results, were also calculated to elaborate on the evidence related to internal structure. The correlation matrix is presented in Table 4. These correlations indicate strong relationships between the sub-domains. However, the sub-domains are distinct from each other; most of the correlations are small to moderate in magnitude.

Table 4

Inter-correlations between Subdomains (SRL Assessment)

Criterion	Sub-criterion	P	M	E	A	MG	SE	MS	MT	HS	ME	U
Metacognition	Planning	-	.74	.66	.27	.42	.41	.23	.46	.39	.32	.64
	Monitoring		-	.78	.20	.43	.41	.24	.42	.37	.28	.63
	Evaluation			-	.14	.41	.34	.24	.44	.35	.29	.56
Motivation	Anxiety				-	.29	.41	.16	.29	.26	.23	.22
	Mastery-goal orientation					-	.52	.26	.39	.31	.27	.46
	Self-efficacy						-	.27	.34	.29	.21	.41
	Mindset							-	.25	.26	.19	.28
Strategies for	Managing time								-	.44	.46	.48
	Help seeking									-	.32	.50
	Managing environment										-	.39
	Understanding											-

NOTE: P = planning, M=monitoring, E = evaluation, A=anxiety, MG=mastery-goal orientation, SE=self-efficacy, MS=mindset, MT=strategies for managing time, HS = strategies for help-seeking, ME= strategies for managing environment, U = strategies for understanding

Reliability. The reliability estimates for this 47-item SRL survey are above .80 at both the domain ($\alpha = .82$) and sub-domain ($\alpha = .83$ to $.84$) levels. Table 5 shows the Cronbach's alphas of the scores for each domain and sub-domain. According to Nunnally (1978), these reliability estimates are acceptable.

Table 5

Central Tendencies and Reliability Estimates by Domains and Sub-domains ($n = 4,614$)

Domains & Sub-domains (# of items)	Cronbach's Alpha
Metacognition (13 items)	.82
Planning (3 items)	.84
Monitoring (6 items)	.83
Evaluation (4 items)	.84
Motivation (20 items)	.82
Anxiety (6 items)	.84
Self-efficacy (4 items)	.84
Mastery-goal orientation (4 items)	.84
Mindset (6 items)	.84
Strategies (14 items)	.82
For managing time (3 items)	.84
For help seeking (3 items)	.84
For managing environment (3 items)	.84
For understanding (5 items)	.84

DAACS Writing Assessment

Content. Construct validity evidence regarding the content of the writing assessment requires the examination of both the writing assessment prompt and the rubric based on which student essays are scored. Both the prompt and the rubric were developed based on best practices and expert advisement. Once completed, drafts of the prompt and the rubric were sent to three experts and ten students for review.

The three content experts evaluated the prompt and rubric in terms of clarity and their alignment to the purpose and each other by responding to five 3-point Likert-type scale items. The specific items and expert responses are provided in Table 6. Results suggest that the prompt

and rubric criteria are generally aligned, and content of the rubric is relevant to the purposes of the writing assessment.

Table 6

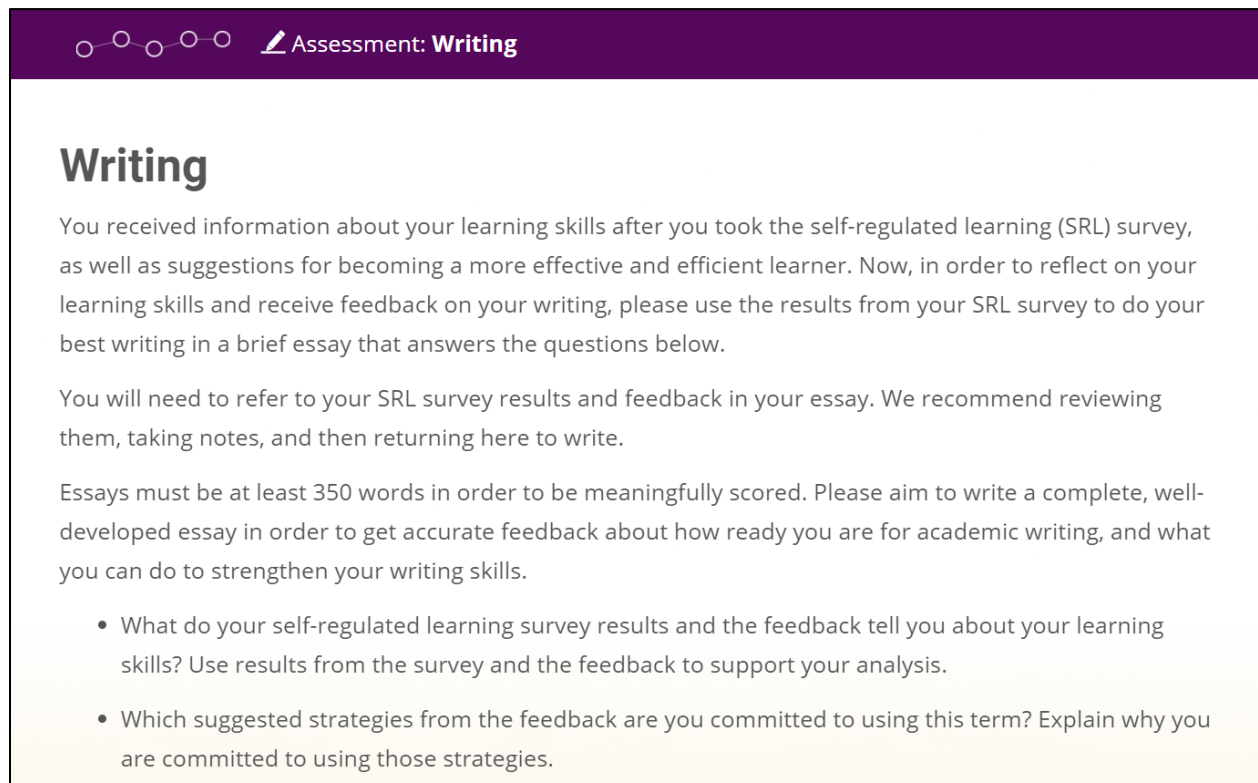
Frequency of Responses from Experts on Statements Evaluating Writing Assessment


Statements Evaluating the Writing Assessment	Yes	Somewhat	No
1. Purposes of writing assessment are aligned with writing prompt	2	1	0
2. Prompt will elicit writing that reflects content of rubric	2	1	0
3. Prompt clearly describes what is expected of students' essays	0	2	1
4. Students might interpret their scores in unanticipated ways	2	1	0
5. Students might have emotional reactions to their feedback that would influence their responses	2	1	0

Ten undergraduate college students were asked to evaluate the prompt and criteria of the writing assessment for clarity, strengths, weaknesses, and suggestions for improvement. General impressions of their responses were positive. For example, one student wrote, "The purpose was clear. It gives students a chance to know exactly what they will need to do." Another wrote, "I think the prompt is very clear about the expectations for the writing assessment. I would not change anything." Criteria were also clear to most of those students, as one pointed out, "It is simple to understand what areas are being evaluated within our writing." Some also expressed appreciation toward the content of the prompt, saying "I like how you can reflect on your own results," which was also echoed by other students.

A couple of students raised concerns about the number of criteria that students have to address in a 500 word essay. For example, one student wrote, "500 words may be too little to cover everything required." Another voiced a similar concern with a different perspective, "A lot of criteria, maybe condense a few of the bullet points?" Students' comments on the weaknesses and suggestions for improvement along with the expert reviews were used for revising the

assessment. The final versions of the prompt and rubric are provided in Figure 4 and Table 8, respectively.

The image is a screenshot of a web interface for a writing assessment. At the top, there is a dark purple header bar. On the left side of this bar is a logo consisting of five white circles connected by lines. To the right of the logo, the text "Assessment: Writing" is displayed in white. Below the header, the main content area has a light beige background. The word "Writing" is prominently displayed in a large, bold, dark grey font. Below this title, there are three paragraphs of text in a standard dark grey font. The first paragraph explains the purpose of the writing task, linking it to a previous SRL survey. The second paragraph provides instructions on how to use the survey results in the essay. The third paragraph specifies the word count requirement and the goal of the assessment. At the bottom of the content area, there is a list of two bullet points, each starting with a dark grey dot. The first bullet point asks about learning skills and feedback, and the second asks about strategies from the feedback. The entire content area is enclosed in a thin black border.

 Assessment: **Writing**

Writing

You received information about your learning skills after you took the self-regulated learning (SRL) survey, as well as suggestions for becoming a more effective and efficient learner. Now, in order to reflect on your learning skills and receive feedback on your writing, please use the results from your SRL survey to do your best writing in a brief essay that answers the questions below.

You will need to refer to your SRL survey results and feedback in your essay. We recommend reviewing them, taking notes, and then returning here to write.

Essays must be at least 350 words in order to be meaningfully scored. Please aim to write a complete, well-developed essay in order to get accurate feedback about how ready you are for academic writing, and what you can do to strengthen your writing skills.

- What do your self-regulated learning survey results and the feedback tell you about your learning skills? Use results from the survey and the feedback to support your analysis.
- Which suggested strategies from the feedback are you committed to using this term? Explain why you are committed to using those strategies.

Figure 4. Revised writing prompt on the DAACS platform.

Table 8

Criteria and Sub-criteria for the Writing Assessment

Criterion	Sub-criterion	Brief Description (as presented to students)
Content	Summary	<ul style="list-style-type: none"> the essay uses the survey results and feedback to create a detailed summary of your strengths and weaknesses as a learner,
	Suggestions	<ul style="list-style-type: none"> contains suggestions you are committed to using, explains your choices of suggestions in terms of your survey results and feedback
Organization	Structure	<ul style="list-style-type: none"> the essay has a clear and logical organization,
	Transitions	<ul style="list-style-type: none"> uses transitions and linking words and phrases to guide readers through the discussion.
Paragraphs	Ideas	<ul style="list-style-type: none"> paragraphs consistently and clearly focus on a main idea or point. sentences are linked together in a way that allows the reader to see the relationship between the ideas or information in one sentence and those in another sentence.
	Cohesion	<ul style="list-style-type: none"> uses adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), and conjunctions (e.g., and, or, while, whereas) to link sentences and ideas together.
Sentence	Correct	<ul style="list-style-type: none"> sentences are correct: no run-ons, fragments, or errors in subject-verb agreement.
	Complexity	<ul style="list-style-type: none"> uses a variety of sentence structures.
Conventions		<ul style="list-style-type: none"> spelling, punctuation, and capitalization are correct.

Internal structure. In order to collect evidence of internal structure for the writing assessment, mean and standard deviation of scores on each sub-criterion, and correlations between them were examined. As shown in Table 9, students generally performed at the medium level, with scores hovering between 2.17 and 2.74 (SD = .46 to .75). Correlations suggest that performance on the sub-criteria are related to each other, with some relationships being stronger than others. However, the sub-criteria are also distinct from one another as most of these correlations are small to moderate.

Table 9

*Descriptive Statistics for and Correlations between Writing Assessment Sub-domains (1=Low, 2=Medium, 3=High)**

Criterion	Sub-criterion	M	SD	SM	SG	SC	TR	ID	CS	CR	CP	CV
Content	Summary (SM)	2.55	0.72	-	.38	.37	.29	.21	.27	.16	.12	.15
	Suggestions (SG)	2.49	0.75		-	.32	.24	.19	.30	.23	.13	.18
Organization	Structure (SC)	2.74	0.46			-	.49	.48	.55	.27	.26	.30
	Transitions (TR)	2.17	0.63				-	.36	.38	.16	.40	.18
Paragraphs	Ideas (ID)	2.83	0.48					-	.47	.28	.12	.31
	Cohesion (CS)	2.78	0.44						-	.28	.24	.29
Sentence	Correct (CR)	2.73	0.49							-	.14	.42
	Complexity (CP)	2.26	0.46								-	.14
	Conventions (CV)	2.73	0.49									-

*Min = 1, Max = 3

Reliability. Percent agreement was used to monitor inter-rater reliability between raters.

Table 10 shows the average and range of percent agreement by criterion for the essays that were scored by two raters, within the first 1000 essays that were scored. As shown, conventions had the lowest average percent agreement (55.33%, SD=8.54), while structure had the highest average percent agreement (62.89%, SD=5.93). The average percent agreement across all criteria ranged between 55.33% and 62.67%. The progression of inter-rater reliability for the first 1000 essays are illustrated in Appendix B by criterion. This information was used to inform the focus of further training and calibration sessions for the raters, which were held by two experts.

Table 10

Percent Agreement by Criterion

Criteria	Mean (SD)	Min	Max
Summary	55.78 (8.69)	36.00	64.00
Suggestions	59.78 (6.00)	51.00	58.00
Structure	62.89 (5.93)	54.00	72.00
Transitions	57.11 (7.90)	45.00	68.00

Ideas	59.78 (5.67)	50.00	68.00
Cohesion	62.67 (9.80)	45.00	72.00
Correct	56.33 (5.05)	48.00	63.00
Complexity	56.00 (6.76)	44.00	68.00
Conventions	55.33 (8.54)	46.00	72.00

Currently, we are also examining the inter-rater reliability between human raters and our automated system, LightSide. These statistics will be reported in our final paper.

DAACS Reading Assessment

Content. Questions and passages for reading assessment were adapted from four years (2011-2014) of New York English Language Arts Regents examinations. A total of 30 passages and 180 items (6 items for each passage) were selected from state-mandated high school English language exams by a content expert. The content of the items was reviewed, and items were subsequently categorized into one of five domains: ideas, inferences, purpose, language, and structure. Difficulty levels and point-biserials were item statistics reported by the state; using this information, the items were also classified into three difficulty levels for the purpose of DAACS: high, medium, low. Table 11 breaks down the 180 items by reading level of passages and content domain.

Table 11

Reading Assessment Items by Passage Reading Level and Content Domain

	Number of passages	Ideas	Inferences	Language	Purpose	Structure	TOT
Easy	13	15	23	27	8	5	78
Medium	7	11	8	7	6	10	42
Hard	10	28	7	8	4	13	60
TOTAL	30	54	38	42	18	28	180

Descriptive Statistics and Internal structure. One point was granted for every item that was answered correctly, and 0 for an incorrect response. Descriptive statistics and reliability were calculated for the overall reading assessment and not by domain because the assessment was designed to be unidimensional. The overall standardized mean score for the reading assessment was 0.89, SD = 0.13 (Min = 0.06, Max = 1.00). Cronbach's alpha was 0.74. Additional analyses of the internal structure are in progress.

DAACS Mathematics Assessment

Content. Two hundred nineteen items were selected from state-mandated high school mathematics exam and reviewed by a content expert. The content of the items was reviewed, and items were subsequently categorized into one of six domains: geometry, lines and functions, number and calculation, representing word problems with algebra, statistics, and variables and equations. Difficulty levels and point-biserials were item statistics reported by the state; using this information, the items were also classified into three difficulty levels for the purpose of DAACS: high, medium, low. Table 12 breaks down the 219 items by item level and content domain.

Table 12

Mathematics Assessment Items by Level and Content Domain

Categories	Easy	Medium	Hard	Total
Geometry	14	11	14	39
Lines and functions	12	13	10	35
Number and Calculation	12	17	13	42
Representing word problems with algebra	11	6	10	27
Statistics	12	11	12	35
Variables and equations	14	13	14	41
Total	75	73	71	219

Descriptive Statistics and Internal Structure. One point was granted for every item that was answered correctly, and 0 for an incorrect response. Descriptive statistics and reliability were calculated for the overall mathematics assessment and not by domain because the assessment was designed to be unidimensional. The overall mean score for the mathematics assessment was 0.62, SD = 0.18 (Min = 0.06, Max = 1.00). Cronbach's alpha was calculated as evidence of internal consistency for the mathematics assessment overall, $\alpha = 0.68$.

Relations between DAACS Instruments and Achievement

Relations to other variables were examined using correlations between the four DAACS assessments – SRL, math, reading, and writing – and within SRL domains. These correlations can be found in Table 13. The predominantly large correlations between metacognition, motivation, strategies, and grit suggest that the SRL survey has strong convergent validity. The low correlations of the SRL survey with the math, reading, and writing assessments also provide strong evidence of their divergent validity, suggesting that the SRL survey is measuring an aspect of student learning and achievement entirely different from what the other assessments measure. The small to moderate correlations between the math, reading, and writing assessments suggest that these assessments are related yet distinct.

Table 13

Convergent and Divergent Validity Evidence

Convergent	1	2	3	4	Divergent	5	6	7	8
1. Metacognition	1				5. SRL Total	1	.07	.08	.04
2. Motivation	.44	1			6. Math Total		1	.36	.24
3. Strategies	.63	.52	1		7. Reading Total			1	.31
4. Grit	.47	.55	.61	1	8. Writing Total				1

Correlations were calculated between each of the four assessments and total number of courses students passed. The correlations with total courses passed are as follows: SRL ($r = .09$), grit ($r = .01$), writing ($r = .10$), reading ($r = .15$), and math ($r = .15$). This indicates that the DAACS reading and math assessments are the most predictive of academic achievement, followed by writing, SRL, and Grit.

Conclusion

As suggested by the framework, evidence of validity for each DAACS assessment (SRL survey, writing, reading, and mathematics) were collected for their content, internal structure, and relationships with each other and achievement. The collective validity evidence is discussed separately for each assessment below.

The SRL survey has strong evidence for its content validity based on the systematic evaluation and selection of relevant items by an expert panel. A panel of experts agreed on a set of criteria for evaluating individual items, which contributed to the content coverage of the SRL survey. Overlap between the content of items was minimized by evaluating individual items based on the importance of what each item measured and whether there were other items measuring the same content.

The internal structure of the SRL survey also provided evidence of the validity of responses. The CFA results confirmed a model based on an EFA; this model reflects three dimensions that overlapped with the cognitive, affective, and behavioral subdomains of SRL.

The writing assessment offers an opportunity to students to reflect on the feedback they received based on the results of the SRL survey. Evidence of the content of the writing assessment was evaluated by having an expert panel evaluate the content of the prompt and the rubric. Overall, experts positively evaluated the instrument, suggesting that the prompt would elicit desired responses and the rubric would be effective in capturing the intended differences in students' writing proficiencies.

Essays are scored at the subcriteria level to provide detailed feedback to students. Evidence of the internal structure of the writing assessment was explored by examining the relationships between scores at the subcriteria level. The correlations between these ratings range from .12 to .49, suggesting that each subcriterion can inform a student on a different aspect of college writing and is still relevant to aspects of writing represented by other subcriteria.

The use of operationally administered assessment items gives creditability to the quality of the items in the reading and math assessments. However, more validity evidence is needed due to the use of computer adaptive testing.

Both the reading and math assessments were designed to be unidimensional. However, more evidence needs to be collected to test the assumption of unidimensionality, and to test the internal structure of both assessments.

Correlations between responses to the SRL survey and scores on the writing, mathematics, and reading assessments provide evidence for construct validity of these assessments in terms of relations between variables. Relationships between these assessments also give an opportunity to understand the interplay between the academic and non-academic aspects of DAACS. The correlations between the non-academic SRL survey and the academic achievement measures (writing, mathematics, and reading) were found to be low. Although almost no linear relationship was reported between any of those instruments, it is worth noting

that the relationships between writing, mathematics, and reading were higher than the relationship between SRL survey responses and scores on those assessments. This suggests that the SRL survey and the academic measures of DAACS are distinct from each other.

Relationships between DAACS instruments and academic achievement in terms of earned course credits were also calculated. Although no linear relationship was found, the correlations with achievement were similar for each of the DAACS assessments. That is, there is a stronger relationship between academic achievement and scores on math and reading assessments than between academic achievement and writing and SRL.

All in all, there is promising evidence of the validity and reliability of the DAACS instruments. Various types of evidence for each assessment informed us about the strengths and weaknesses of each of the DAACS assessments. Each instrument provides a snapshot of students' preparedness for college. Given the diagnostic purpose of these assessments, further validity evidence is being collected on the accuracy of the portrayal of the student strengths' and weaknesses and the overlap between students' needs and the feedback.

Educational Implications

DAACS has the potential to reduce the need for remedial courses by providing personalized and efficient guidance for students. DAACS instruments can be used for not only understanding students' level of preparedness, but also for providing resources to the student at no cost and inline with their regular course taking. It is one of the first efforts to provide non-academic and academic support to students in order to address the retention problem in higher education.

References

- AERA, APA, & NCME (2014). *Standards for educational and psychological testing* (8th ed). Washington, DC: American Educational Research Association.
- Bailey, T., & Cho, S. W. (2010). Issue Brief: Developmental Education in Community Colleges. *Community College Research Center, Columbia University*.
- Belfield, C. R., & Crosta, P. M. (2012). Predicting Success in College: The Importance of Placement Tests and High School Transcripts. CCRC Working Paper No. 42. *Community College Research Center, Columbia University*.
- Donker, A., de Boer, H., Kostons, D., Dignath van Ewijk, C., & van der Werf, M. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review 11*, 1–26.
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), 1087.
- Lai, E. (2011). Metacognition: A literature review. Retrieved from http://www.pearsonassessments.com/hai/images/tmrs/Metacognition_Literature_Review_Final.pdf
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Scott-Clayton, J. (2012). Do High-Stakes Placement Exams Predict College Success? CCRC Working Paper No. 41. *Community College Research Center, Columbia University*.

Valentine, J. C., Konstantopoulos, S., & Goldrick-Rab, S. (2017). What happens to students placed into developmental education? A meta-analysis of regression discontinuity studies. *Review Of Educational Research*, 87(4), 806-833.

Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. In B. J Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance: Educational Psychology Handbook*. New York: Routledge.

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and overview. In B. J Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance: Educational Psychology Handbook*. New York: Routledge.

Appendices

Appendix A: DAACS SRL Survey

Item Name	Actual Survey Item
anxiety1	During important exams, I think that I am doing awful or that I may fail.
anxiety2	I feel out of sorts or not really myself when I take important exams.
anxiety3	During important exams, I cannot remember material that I knew before the exam.
anxiety4	The closer I am to a major exam, the harder it is for me to concentrate on the material.
anxiety5	When I study for my exams, I worry that I will not remember the material on the exam.
anxiety6	I worry so much before a major exam that I am too worn out to do my best on the exam.
evaluation1	I ask myself if I learned as much as I could have once I finish a task.
evaluation2	I ask myself how well I accomplished my goals once I'm finished.
evaluation3	I summarize what I've learned after I finish.
evaluation4	I ask myself if I have considered all options after I solve a problem.
help_seeking1	I ask others for help when I don't understand something.
help_seeking2	I avoid asking questions about things I don't understand.
help_seeking3	I ask my instructor questions when I do not understand something.
managing_environment1	I make sure no one disturbs me when I study.
managing_environment2	I try to study in a place that has no distractions (e.g., noise, people talking).
managing_environment3	I let people interrupt me when I am studying.
managing_time1	I wait to the last minute to start studying for upcoming tests.
managing_time2	I pace myself while learning in order to have enough time.
managing_time3	I finish all of my schoolwork before I do anything else.
mastery_orientation1	I find coursework enjoyable.
mastery_orientation2	I want to master the things I am learning.
mastery_orientation3	What I am learning is relevant to my life.
mastery_orientation4	Learning is fun for me.
mindset1	You have a certain amount of intelligence, and you can't really do much about it.
mindset2	No matter who you are, you can significantly change your intelligence level.
mindset3	You can always greatly change how intelligent you are.
mindset4	Your intelligence is something about you that you can't change very much.
mindset5	You can learn new things, but you can't really change your basic intelligence.
mindset6	No matter how much intelligence you have, you can always change it quite a bit.
monitoring1	I ask myself periodically if I am meeting my goals.
monitoring2	I find myself analyzing the usefulness of strategies while I study.
monitoring3	I ask myself questions about how well I am doing while I am learning something new.
monitoring4	I consider several alternatives to a problem before I answer.
monitoring5	I find myself pausing regularly to check my comprehension.
monitoring6	I ask myself if what I'm reading is related to what I already know.
planning1	I think of several ways to solve a problem and choose the best one.

planning2	I think about what I really need to learn before I begin a task.
planning3	I ask myself questions about the material before I begin.
self_efficacy1	I am confident I can learn without the physical presence of an instructor to assist me.
self_efficacy2	I am certain I can understand even the most difficult material presented in an online course.
self_efficacy3	I am confident I can do an outstanding job on the activities in an online course.
self_efficacy4	Even with distractions, I am confident I can learn material presented online.
understanding1	I consciously focus my attention on important information.
understanding2	I stop and go back over new information that is not clear.
understanding3	I think about the types of questions that might be on a test.
understanding4	I stop and reread when I get confused.
understanding5	I make pictures or diagrams to help me learn concepts.

Appendix B

Percent Agreement between Raters for Writing Assessment (first 1000 essays)

