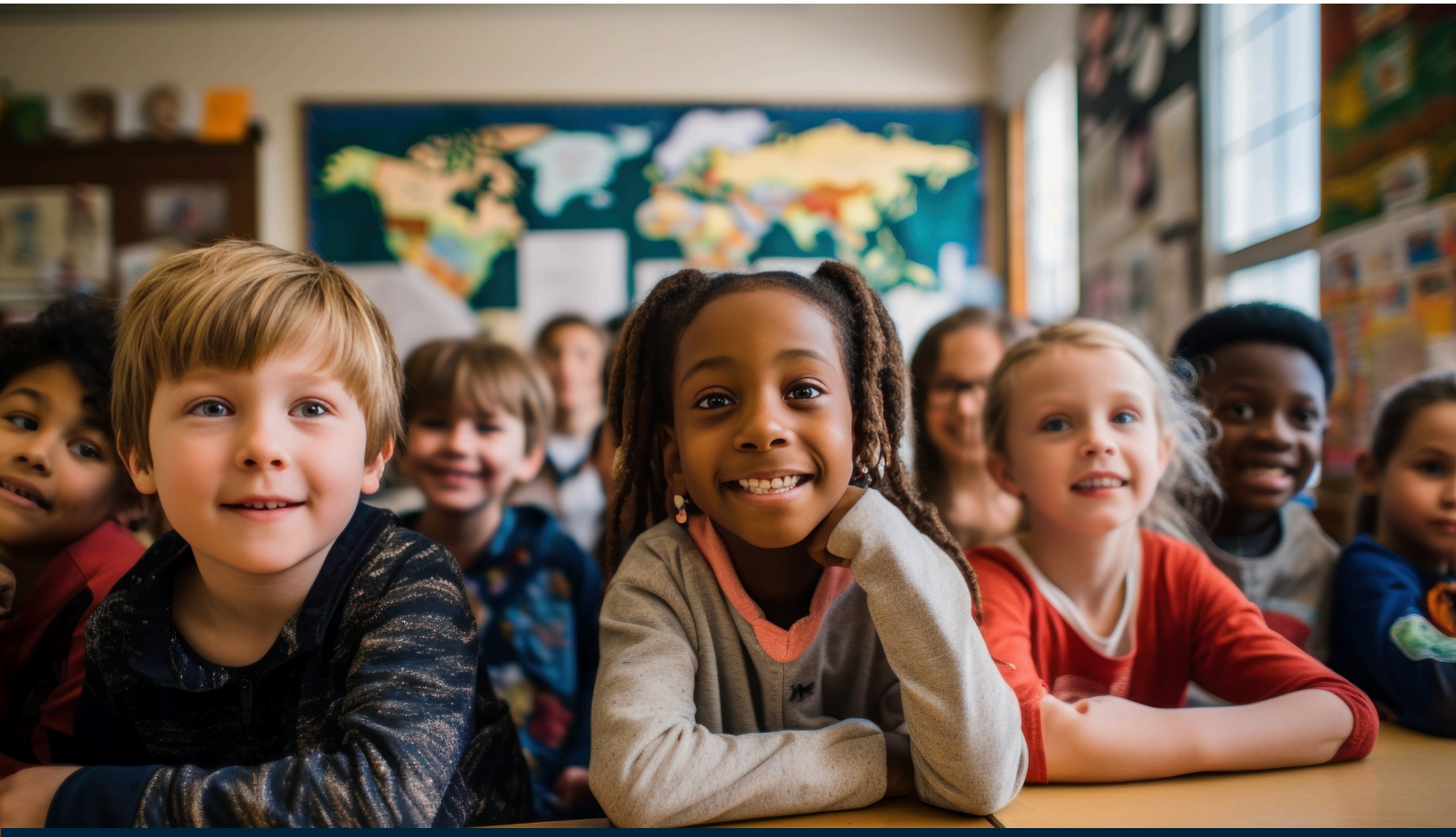
Emma Eccles Jones College of Education & Human Services
Center for the School of the Future
UtahStateUniversity

SB173 Teacher Merit Award: Developing Reliable and Valid Assessments for Student Growth in State Non-Assessed Subjects

A Policy and Research Brief

April 2025



The purpose of the Center for the School of the Future is to promote empirically validated practices in public education systems and to encourage cooperative and research relationships between K-12 and higher education institutions.

The Context

The purpose of Senate Bill 173 and 99, the Teacher Merit Awards (TMA), is to pilot a program that recognizes the top 25% of Utah's teachers through a sizable monetary award given for three consecutive years. The Center for the School of the Future (CSF) at Utah State University is responsible for the administration of this pilot program.

The pilot program highly values the ability of top-performing teachers to advance student learning. Thus, Local Education Agencies (LEAs) must determine which teachers produce student learning over and above what can typically be expected of each student. If students in a teacher's classroom score above what would typically be expected of them, then we can say that teacher "adds value" to student learning. The concept of "value-added" is central to identifying top-performing teachers. It highlights the extent to which a teacher's instruction fosters student learning beyond typical expectations, providing a fair and growth-focused measure of teacher effectiveness.

For the purposes of SB 173, teachers' "value-added" will be determined by state-mandated assessments in English language arts, math, and science. We recognize that state mandated assessments are not available for all subject areas taught in Utah secondary schools. The purpose of this policy and research brief assists teacher teams in the design and development of **reliable** and **valid** instruments to assess student growth in state non-assessed subjects to participate in SB173 and 99, Teacher Merit Award. The brief addresses written assessments only; it does not address the performance arts or PE where performances, products, or projects are used to evaluate student growth.

At the end of this brief, your team will be able to describe the steps necessary to:

1) develop a written assessment to determine student growth in a subject area not assessed by the state

2) identify the appropriate type of evidence your team will use to ensure the **reliability** of your team's written assessment of student growth; and

3) identify the appropriate type of evidence your team will use to ensure the **validity** of your team's written assessment of student growth

Developing Written Assessments in State Non-Assessed Subjects

To determine student growth in most secondary school subject areas, LEAs will need to find and/or develop student growth assessments. To effectively develop assessments in all subject areas taught, LEAs should establish collaborative teams that identify critical standards, review existing tools, and pilot assessments. Regular coordination meetings, scheduled bi-weekly or monthly, can help maintain alignment and accountability throughout the process. LEAs are encouraged to work together collaboratively to search district websites (e.g., Austin, Texas; Delaware; Hillsborough County, Florida), professional organizations (e.g.,

National Council of Social Studies; Association for Career and Technical Education), and other resources to find psychometrically sound assessments for non-state assessed subject areas. The National Assessment of Educational Progress' (NAEP) website contains example questions, not only for math, reading, and science, but also for civics, U.S. history, geography, economics, and more. Well-developed items can also be found in examples from national assessments, such as Smarter Balanced, The Partnership for Assessment of Readiness for College and Careers (PARCC), and Utah's ASPIRE and Readiness, Improvement, Success, and Empowerment (RISE). Some examples of items from these assessments can be found on the Internet. Other professional organizations often provide information and examples related to assessments in their disciplines as well.

If non-assessed subject area assessments cannot be located or are psychometrically unacceptable, subject area teacher teams can develop their own assessments, so long as the assessments 1) are administered at two points in time, and 2) provide adequate reliability and validity evidence approved by CSF, and 3) are able to determine average growth and value-added growth based on the results of the assessment.

Start With Knowledge Domains and Content Discipline Standards

Most subject area knowledge domains have an accepted knowledge base and/or a set of content discipline standards provided by the state or by professional organizations. In addition, most teachers set up learning objectives ahead of time for their subject area content. Disciplinary or content area standards and objectives are a good place to start when developing an assessment for a subject area. Consider some of the following questions as your team gets started: What does your team want students to know and be able to do after taking the course? What competencies or skills will they demonstrate? What concepts are critically important in your team's subject area?

- Work with colleagues in your team's subject area to determine the critical concepts and competencies/skills your team wants to measure. This is best done in a small group of professionals with similar backgrounds.
- Most Utah core standards have a general statement followed by a set of sub-standards. Begin with the general standards.

Develop Assessment Items from the Knowledge Domain Specifications and Content Disciplinary Standards

Once the content domain knowledge, e.g., U.S. History, has been clearly and carefully specified, your team will need to identify specific learning objectives around the critical concepts, standards, or competencies/skills within that domain of knowledge or associated with that domain's accepted standards. These learning objectives will form the base for developing assessment items.

Your team can develop assessment items on your own or use ChatGPT to develop them. We recommend using ChatGPT as it will save time and energy in the development of your team's assessment.

1) Here is how to use ChatGPT to develop assessment items based on your team's

content standards and/or learning objectives Visit https://chatgpt.com/ and create an account to use a free ChatGPT service.

2) Define Content Area(s) –

- You can use different formats to prompt ChatGPT to generate test questions, including:
  a. *Content area standards*
  b. *Learning objective.*
  c. *A phrase or sentence about the content area(s)*
  d. *The visuals include graphs, charts, data tables, images containing text, and videos*
- Provide context for ChatGPT, including grade level, students' accomplishment levels from previous grades, any applicable disabilities, and the learning characteristics of the student population in a class or school.

3) Generate Initial Assessment Items – Provide ChatGPT with specific prompts based on the learning objectives. For example:

- "*Generate five multiple-choice questions assessing students' understanding of the causes of the American Revolution, ensuring that each question aligns with Bloom's Taxonomy*."

4) Review and Revise Questions – Ensure that the AI-generated items align with the intended depth of knowledge, follow best practices for assessment design, and avoid bias or ambiguity.

5) Pilot Test Items – Before full implementation, test the questions with a small group of students or colleagues to identify any inconsistencies or misinterpretations.

6) Analyze Item Performance – Use item analysis techniques, such as difficulty index and discrimination index, to refine the questions based on student responses.

7) Finalize and Integrate Questions – After revisions, incorporate the validated items into the assessment framework, ensuring alignment with instructional goals.

Considerations for Using ChatGPT

- **AI as an Assistant, Not a Replacement** – ChatGPT should be viewed as a support tool, not a substitute for professional judgment in assessment design.

- **Validity and Reliability** – ChatGPT-generated questions should undergo expert review and statistical validation to ensure they accurately measure student learning.

- **Alignment with Standards** – Educators must verify that the AI-generated questions align with state standards, curriculum frameworks, and instructional objectives.

Once assessment items are gleaned from ChatGPT, your team will need to edit each item to ensure that it matches each learning objective and meets additional recommendations from Burton (1991) for developing effective multiple-choice assessment items:

Each item should assess no more than one learning objective or standard.

- Base each item on a specific problem stated clearly in the stem.
- Include as much information of the item in the stem as needed, but do not include

irrelevant information.
- State the stem in positive, not negative form.
- Word the alternatives clearly and concisely.
- Keep the alternatives mutually exclusive.
-
- Keep the alternatives homogenous in content.
- Keep the grammar of each alternative consistent with the stem.
- Keep the alternatives parallel in form.
- Keep the alternatives similar in length.
- Avoid textbook, verbatim phrasing.
- Avoid the terms always, never, and only.
- Avoid using keywords in the alternatives.
- Use plausible distractors.
- Avoid the phrases “all of the above” or “none of the above.”
- Include one and only one clear and correct answer. Avoid ambiguous
- alternatives.
- Present the correct answer within the alternatives in random order.
- Avoid unnecessary difficult vocabulary.
- Analyze the effectiveness of each item.

Be aware that the items generated from ChatGPT may not be consistent with Burton’s (1991) recommendations. Burton’s recommendations should be used as the standard rather than ChatGPT. Consult https://assessmenting.byu.edu/handbooks/betteritems.pdf for specific examples of poor and better examples as your team develops items.

Assessment developers suggest a minimum of 4-6 items for each learning objective based on your team’s established or stated concepts, standards, and competencies/skills. It is important to develop more items than your team needs, since some will be eliminated after feedback from colleagues and students in pilot trials and after reliability analyses, such as item-intercorrelations/item response theory (IRTs).

Once your team has developed enough assessment items, pilot those items with other subject area teachers and ensure that all agree on the correct responses. Each assessment item should stand on its own; in other words, items should not have to be explained verbally to produce the correct response. If subject area teachers disagree, then the item needs to be either modified or eliminated. Once teachers have reviewed the items, send them out for content knowledge expert review (e.g., district specialists, university faculty). If experts do not agree on an item, modify or eliminate the item.

Pilot and Assessment items

Once the items have been developed, administer them to a few students who resemble the students you want to respond to this assessment. Students’ feedback might point to items that are confusing, items they cannot understand because of unfamiliar vocabulary, items that might have two possible and reasonable responses. Change or delete items based on feedback from these students’ scores and on any focus-group sessions following your administration.

Develop a Standard Administration Protocol

Develop written instructions for students taking the assessment. All subject area instructors who administer the assessment will use the same instructions script. Also, identify a specific window of time in which the assessment will be administered and the length of time allowed for students to respond to the assessment. Assessments should be administered as a pre-assessment

I at the beginning of class/course and as a post-assessment at the end of class/course to measure student growth.

Develop a Standard Scoring Protocol

It is recommended that the assessment be administered online using Survey Monkey, Qualtrics, or another similar platform. The value of administering assessments online is twofold: 1) online assessments are easier to score, and 2) many platforms (i.e. Qualtrics) include descriptive statistics that will be useful in producing summative or formative reports, such as totals, standard deviations, percentages correct, correlations among items, and more.

Establishing Reliability and Validity of Written Assessments

Reliability and validity of scores are the foundation of any psychometrically sound assessment to be administered as part of SB173 participation. If LEAs want to include student growth assessments in non-state assessed subject areas, they will be responsible for providing evidence that their assessments are reliable and valid.

Written Assessment Reliability

Once assessment items for each standard, concept, or competency/skill for your subject area have been piloted and revised, the next step is to establish score reliability. Reliability refers to the consistency of an assessment. For example, a reliable test produces similar results when administered multiple times under the same conditions. Assessments must evidence reliable scores as a prerequisite for them to be valid. To establish the reliability of an assessment, you need to answer the question: *Does the assessment measure concepts, standards, and competencies/skills* in the subject area consistently and accurately?

A good way to think about reliability is to compare it to shooting at a target. Think of the target as the concepts, standards, and competencies or skills your team wants to assess. Think of the arrows as specific items your team has developed. Each arrow (individual items) shot at a target should come as close to the bullseye in the target as possible. In other words, each arrow should hit the target accurately and skillfully. When all the arrows hit the target and in the same area, we would say they “correlate” or relate well to one another. When this happens in an assessment, we say the assessment is reliable.

The concept of correlation is important for understanding the reliability of written assessments. If students’ scores on each assessment item perfectly relate to the other assessment items within the written assessment, then the “correlation coefficient” of the assessment would be 1.0. While it is expected that the assessment items relate to one another within the assessment, perfection is not achievable since the items are not the same. An acceptably reliable assessment should have a correlation coefficient of .80 or above. Standardized commercial assessments typically have reliability correlation coefficients of .90+.

Table 1
Types of Reliability for Written Assessments

TYPES OF RELIABILITY FOR WRITTEN ASSESSMENTS		
Internal Consistency	**Score Stability**	**Equivalent, Parallel, or Alternate Forms**
How well are assessment items consistent with one another? Items included in an assessment should include only the concepts, skills, or knowledge taught in the target course and exclude all others. A math test on fractions should include items that consistently assess different aspects of fraction operations, ensuring alignment within the content domain. Internal consistency is determined statistically in three different ways: split-half —odd numbered items are compared to even numbered items and a Spearman-Brown formula is applied to the correlation between the two; *Kuder-Richardson* (K-R)compares all items to one another in all possible ways and a correlation coefficient is produced (K-R 20 assumes all questions are equally difficult, and K-R 21 assumes they are not); and *Cronbach's alpha*, often used when item responses are not binary choices, e.g., yes or no, true or false.	*How stable are scores from one administration to another?* A reliable assessment should have stable scores that do not vary over time. Students scoring similarly on a vocabulary quiz administered on two different days demonstrates the assessment's stability over time. If students are administered an assessment on Day 1, their scores should be similar to their scores if the assessment were administered on Day 3. A Pearson Product Moment correlation coefficient (*r*) is the typical statistical choice for providing evidence of score stability. Typically, test stability scores are reported using a Pearson Product Moment (*r*) correlation coefficient. This correlation coefficient should be $r = .70+$ or higher.	*How well does one version of an assessment relate to another version of the assessment measuring the same concepts, skills, or knowledge?* If two versions of an assessment (e.g., Form A and Form B) are developed, they should be highly correlated. Typically, alternate forms of commercial assessments evidence a Pearson Product Moment *(r)* correlation coefficient of $r = .90+$.

Since it is unlikely that your team will develop alternate forms of your team's assessment, your team should focus on determining internal consistency and score stability as evidence of reliability. Internal consistency is generally determined through examination of item intercorrelation coefficients. Internal consistency will increase when the assessment includes only items that measure concepts, skills, or knowledge taught in the course or expected by the standards.

The following website presents a useful discussion and assistance in using statistics to provide evidence of internal consistency. Under the subheading "Cronbach's Alpha," you can find a link to a spreadsheet you can use to statistically assess internal consistency. This spreadsheet will calculate Spearman-Brown, Kuder-Richardson, and other statistical methods for evaluating the reliability of written assessments. Visit https://researchbasics.education.uconn.edu/instrument reliability/ to access these resources.

Establishing written assessment score stability is often done using a test/retest procedure. In a test-retest process, the written assessment is given to group of students, and then readministered within a brief time, usually at least one day later but not more than a few days. Students should score similarly across both assessment occasions, indicating that the assessment provides stable scores across assessment occasions. A stable score correlation coefficient represents the score of students on the assessment between the first and second assessment administration occasions. However, since it is highly unlikely that students will score the same across both occasions (resulting in a score correlation coefficient of 1.0), test-retest correlations should be .70+ or higher. Such scores would represent score stability.

Written Assessment Validity

Assessment validity is a demonstration that the assessment measures what it claims to measure. For example, until recently, writing assessments often did not include the actual production of writing but instead measured knowledge of related skills, such as grammar, spelling, and/or punctuation—typical editing skills. While these editing skills have been shown to be correlated with actual writing achievement, they do not directly measure students' abilities to generate more or better writing. Thus, from a measurement perspective, these writing assessments are not valid assessments of writing. A valid assessment of writing would necessarily include a measure of actual writing production.

An assessment that has been demonstrated to be reliable is not necessarily valid. Your team may be familiar with many writing assessments that reliably measure grammar, spelling and so forth, but as your team can see, they are not valid. Our target example may be illustrative. Arrows that group together on the wrong target, say, the neighbor's fence, may be reliable because they all hit the neighbor's fence, but they aren't valid without hitting the intended target. What good is achieved when we hit the wrong target, even if most of the arrows hit it?

Once an instrument shows score reliability, then establishing validity is the next step in the evaluation of a psychometrically sound

assessment. There are different types of written assessment validity. Table 2 provides descriptions and examples of different types of assessment validity. To assert an assessment is psychometrically sound, your team must be able to answer such questions as: *Does the assessment measure the same content as what has been taught in my class (content validity)? Is the assessment able to predict student success in a health course (predictive validity)?*

Content validity can be determined through an examination of the items in an assessment compared to items in similar established commercial or national assessments. Items from these assessments can be compared to items your team has written to establish the content validity of your team's assessment. Your team can also examine professional standards in your team's subject area and compare them to the standards used to generate or write assessment items. Finally, your team's assessment should demonstrate predictive validity. Students who score high on the assessment should be the students with the highest grades in your team's course. Similarly, students who score low on the assessment should be the students with the lowest grades in your team's course. The following website presents a useful discussion of validity: https://researchbasics.education.uconn.edu/instrument_validity/

Table 2
Types of Validity for Written Assessments - Titles and Definitions/Examples

Content Validity

Does the assessment measure the same content/domain knowledge or standards as what has been taught? This is typically determined by expert judgment. An assessment measuring student learning in a civics course should include items that measure the civics topics covered in that course.

Concurrent Validity

Does the assessment you've developed resemble other existing, established assessments that claim to be measuring the same concepts, e.g., the NAEP Civics examination? Can you, the assessment developer, point to another assessment that has already been established as valid? A civics assessment should be like other established civics assessments.

Predictive Validity

Can the assessment accurately predict later achievement or student growth? Can students' scores on the assessment predict students' grades at the end of the course? For an assessment to have predictive validity, those students who receive grades of A or B should be the same students who score well on a post-assessment, and students who receive a C or D should be the same students who score poorly on a post-assessment.

Face Validity

Does the assessment appear or “look like” it measures what it says it’s measuring? Some assessments do not appear to measure a subject area, even though the assessment purports to measure that subject area. A math assessment to assess problem solving that includes items that assesses vocabulary knowledge, like commutative property and divisor, cannot be said to have face validity.

Construct Validity

Does an assessment adequately measure changes in an underlying theoretical construct that one wants to measure? Has the construct been accepted to be theoretically meaningful in the field or discipline? Construct validity is typically established using either exploratory or confirmatory factor analyses, a sophisticated statistical examination of assessment items and how they cluster around factors that can be seen to connect to the knowledge domains the assessment items were intended to measure.

Consequential Validity

Does the assessment have unintended positive or negative social consequences? For example, many people might consider reading assessments that result in third graders being retained to have negative consequences. Assessments of academic or social interventions can have positive or negative consequences for continuation of those intervention programs.

Conclusion

Developing reliable and valid assessments is important any time educators wish to develop assessments that will accurately measure student learning. All major commercial assessments were years in development, testing, and revision and often cost millions of dollars to produce. For the purposes of SB 173, available state assessments can and should be used to determine student growth in those state-tested subject areas. However, together with a group of peers, it is possible for teachers to develop reliable and valid assessments to determine student growth for subject areas not assessed by the state mandated assessments.

Table 3 provides a helpful checklist for determining whether an assessment is both valid and reliable.

Table 3
Checklist for the Reliability and Validity of an Assessment

Checklist for the Reliable and Validity of an Assessment

- Do the assessment items follow closely the recommendations of Burton and colleagues (1991)?
- Have you established an acceptable reliability for the assessment?
- Does the assessment measure the content of my course?
- Does the assessment measure the concepts and content standards that are most important in my course?
- Does the assessment exclude irrelevant content?
- Does the assessment compare closely with other standardized assessments that measure similar concepts and content standards?
- Does the assessment accurately predict which students are most successful in the course?

Determining Value-Added Growth Using Pre/Post Assessments

Finally, once teacher teams have established their assessment's reliability and validity, the last step in the development of pre/post assessments for any subject area is to determine how teacher teams will evaluate value-added growth. To evaluate such growth, a minimum of two data points are necessary, preferably three.

References

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice assessment items: Guidelines for university faculty*. Brigham Young University Assessment Services and the Department of Instructional Science. https://assessmenting.byu.edu/handbooks/betteritems.pdf.

Gay, L. R. (1985). *Educational evaluation and measurement: Competencies for analysis and application* (2nd ed.). Charles E. Merrill Publishing.

Kubiszyn, T. & Borich, G. D. (2024). *Educational testing and measurement* (12th ed.). Wiley.

Pressley, M. & McCormick, C. B. (1995). *Advanced educational psychology for educators, researchers and policymakers*. Harper-Collins.

Acknowledgment

This policy brief was supported by funding from the Center for the School of the Future at Utah State University.

Parker Fawson
Director
Center for the School of the Future
Emma Eccles Jones Endowed Chair in Early Education

David E. Forbush
Associate Director
Center for the School of the Future

John Jeon
Learning Scientist
Center for the School of the Future

D. Ray Reutzel
Distinguished Research Fellow
Center for the School of the Future

Janice A. Dole
Distinguished Research Fellow
Center for the School of the Future

Reid Newey
Senior Fellow
Center for the School of the Future

Sam Jarman
Senior Fellow
Center for the School of the Future

Nissa Boman
Assistant
Center for the School of the Future

Kalie Chamberlain
Graduate Assistant
Center for the School of the Future

Jed Grunig
Graduate Assistant
Center for the School of the Future