The AIR Professional File

ARTICLE 144

Association for Institutional Research

UNIVERSITY-WIDE ASSESSMENT DAYS: THE JAMES MADISON UNIVERSITY MODEL

Dena A. Pastor, Kelly J. Foelber, Jessica N. Jacovidis, Keston H. Fulcher, Derek C. Sauder, and Paula D. Love

About the Authors

The authors are with the Center for Assessment and Research Studies, James Madison University.

Abstract

James Madison University has used dedicated Assessment Days for more than 30 years to collect longitudinal data on student learning outcomes. Our model ensures all incoming students are tested twice: once before beginning classes and again after accumulating 45–70 credit hours. Although each student completes only four instruments during a 2-hour testing period, 25 different assessments are administered, thereby allowing for the examination of student growth on a variety of different outcomes. This article describes our model and outlines the logistics involved in planning for Assessment Day, including the physical and human resources needed for its success. We also address changes we have made over the years and the challenges we continue to encounter. Our intention is to share lessons learned and encourage readers to consider how our model might be adapted for the assessment of programs both large and small at their own institutions.

Keywords: Assessment Days, large-scale assessment, general education assessment, data collection designs

Background

Every campus has wide-reaching programs intended to affect the learning and development of all or most students. Examples include general education, large-scale student affairs programs, and campus-wide initiatives. Given the large number of students served by these programs and the importance of their associated outcomes, the effectiveness of these programs is often of great interest to many stakeholders. Assessment data are therefore collected to reveal the strengths and weaknesses of these wide-reaching programs, and to partially fulfill requirements of accrediting bodies and funding agencies. Procuring assessment data that help universities improve student learning and demonstrate accountability, however, is no trivial task. To acquire meaningful information, colleges must carefully consider the data collection design along with the numerous other details inherent in conducting quality research.

The purpose of this article is to describe the approach James Madison University (JMU) has used for more than 30 years to collect assessment data for its university-wide programs. Like other universities, we use dedicated Assessment Days (Swing, 2001). Our Assessment Day approach enables the university to collect longitudinal data on student learning and developmental outcomes by setting aside 2 days per year dedicated to assessment. All incoming first-year students (excluding transfer students) are required to participate in Fall Assessment Day ($N \approx 4{,}000$ students); all students with 45–70 credit hours (typically sophomores and including transfer students) are required to participate in Spring Assessment Day ($N \approx 4{,}000$ students). During Spring Assessment Day students are administered the same instruments they were administered during Fall Assessment Day (18 months prior), thereby creating a pretest–posttest design that permits evaluation of gains in student learning and development.

Before describing Assessment Day logistics and resources, it is important to explain the two primary reasons why we've used this model for more than 30 years. First, our Assessment Day model addresses major weaknesses

associated with common assessment approaches, specifically those using a posttest-only design, cross-sectional data, or convenience samples. Second, we continue to use the Assessment Day model because it allows many questions about student learning and development to be addressed. We provide several examples below to convey the methodological advantages of our approach, the kinds of questions that can be addressed, and how the results are used.

One of the greatest strengths of our Assessment Day model is the assessment of all incoming first-year students the week before classes begin. Results from Fall Assessment Days are used to explore the appropriateness of allowing course credit for various precollege experiences, as illustrated with the results in Table 1 for the American Experience assessment, which is used to assess our American History and Political Science requirement. The similar performance of incoming first-year students with and without dual-enrollment transfer credit on this and many of our other assessments has fueled a continuous debate at our university as to whether dual-enrollment credit should be permitted.

Most importantly, Fall Assessment Day results allow for a richer and more-nuanced interpretation of Spring Assessment Day results. To illustrate, Table 2 provides the percentage of students meeting the faculty-set standard on a quantitative and scientific reasoning assessment at pretest (Fall Assessment Day) and at posttest (Spring Assessment Day). Because a larger percentage of students met the standard at posttest than at pretest, we can conclude that students are gaining in knowledge over time. If we had only posttest data, it could be argued that the posttest results reflect nothing more than the knowledge students had upon arriving at the university. Thus, Fall Assessment Day results allow us to explore—and often rule out—a plausible and competing alternative hypothesis for the posttest findings.

Table 1. Number Correct Mean and Standard Deviation on the 40-item American Experience Assessment for Incoming First-Year Students (N = 925) in 2017 by Type of Course Credit

Type of Course Credit	N	M	SD
Advanced Placement	57	29.4	5.7
Dual Enrollment	71	21.6	5.7
None	797	21.8	6.1

Table 2. Percentage of Students Meeting Standard on Quantitative and Scientific Reasoning Assessment on Fall and Spring Assessment Days for Two Cohorts

N	Fall Assessment Day (Pretest) Year	%	Spring Assessment Day (Posttest) Year	%
367	2015	21%	2017	46%
412	2016	28%	2018	39%

By having each student complete the same assessment twice during the first 18 months of their college career, we are also able to provide evidence of student learning. To illustrate, effect sizes capturing the number of standard deviation units by which average scores change from Fall to Spring Assessment Day are provided in Table 3 for assessments administered to incoming first-year students in 2014. The effect sizes are positive, which indicates that the college experience adds value. The fact that some of the effect sizes are not as large as we would like them to be is a call to action. For example, when the quantitative and scientific reasoning test results indicated that students who had completed their requirement were still struggling to discriminate between correlation and causation, the program director organized a series of faculty meetings to identify student misconceptions and design learning strategies to implement new pedagogies.

Given that the credit window for Spring Assessment Day captures students at various stages of general education completion, our pretest–posttest design also allows change over time to be explored for different subsets of students (Pieper, Fulcher,

Table 3. Effect Sizes for Six Assessments for Students Tested on Fall Assessment Day, 2014, and Spring Assessment Day, 2016

Acronym	Test Name	Content Area	N	d
NW9	Natural World—version 9	Quantitative & scientific reasoning	194	0.53
GLEX2	The Global Experience—version 2	Global history & issues	243	0.37
AMEX3	The American Experience—version 3	American history & political science	246	0.33
ISNW-A1	Institute for Stewardship of the Natural World—version A1	Environmental stewardship	413	0.40
KWH8	Knowledge of Wellness and Health—version 7	Wellness & health	253	1.33
SDA-7	Sociocultural Domain Assessment—version 7	Sociocultural understanding	295	0.77

Note. Effect sizes (d) were calculated by subtracting the Fall 2014 average score from the Spring 2016 average score and dividing by the Fall 2014 standard deviation. The d values can be interpreted as the number of standard deviation units by which the Spring 2016 average differs from the Fall 2014 average. With the exception of the ISNW-A1, results are based on only those students who had completed their content area requirement through coursework at our university by Spring 2016. Because there is no such requirement in environmental stewardship, results for the ISNW-A1 are based on all students who completed the test in both fall and spring.

Sundre, & Erwin, 2008). For instance, students who have yet to take any courses in a general education program are compared to students who have partially completed or fully completed the program (as shown for our American History and Political Science requirement in Table 4.[1] (See also Hathcoat, Sundre, & Johnston, 2015, Tables 6 and 7.) Furthermore, we consider score differences among students who have completed their requirements elsewhere (e.g., transfer credits, Advanced Placement credits), allowing us to explore the impact of non-JMU coursework.[2]

Because of the advantages of our Assessment Day model, we continue to use it year after year. Of course, the current design looks quite different from how it looked 30 years ago. In response to challenges encountered along the way, many modifications have been made—and continue to be made—to our Assessment Day model. In the sections below, we build on the work of Grays and Sundre (2012) by describing our model and sharing what we have learned from its implementation. Specifically, we detail the logistics involved, highlighting physical materials and communication strategies. We also describe the logistics team and its responsibilities before, during, and after Assessment Day. Furthermore, we describe the important role that proctors play on Assessment Day and the process we use for their hiring and training. The paper concludes with discussion of changes we have made to Assessment Day and the challenges we continue to encounter.

[1] The results in Table 4 are typical of the kind of results we see on many of our assessments. We often see gains in knowledge over time, but not of the magnitude we would like. As well, increased coursework in the domain is often not strongly related to pretest–posttest gains. Faculty reactions and explanations for such results are provided in Mathers, Finney, and Hathcoat (2018).

[2] Of course, because students were not assigned randomly to these different experiences we cannot claim that different kinds or amounts of coursework cause these score changes. To strengthen the causal link between assessment results and experiences we've used alternative analytical techniques (e.g., propensity score analysis; Harris & Horst, 2016) and implementation fidelity studies, which consider the extent to which programs are delivered as intended (Fisher, Smith, Finney, & Pinder, 2014; Gerstner & Finney, 2013; Swain, Finney, & Gerstner, 2013).

Table 4. Number Correct Mean (and Standard Deviation) on the 40-item American Experience Assessment on Fall and Spring Assessment Days by Course Completion Status

	N	Fall Assessment Day (Pretest) 2016	Spring Assessment Day (Posttest) 2018
JMU course completed			
American History	150	23.1 (5.3)	25.3 (5.6)
Political Science	71	24.1 (5.6)	25.0 (5.4)
JMU course not completed			
Not currently enrolled in American History/Political Science course	85	22.7 (5.9)	23.1 (5.7)
Currently enrolled in American History/Political Science course	52	21.6 (5.9)	23.7 (5.4)

Note. These results are typical of the kind of results we see on many of our assessments. We often see gains in knowledge over time, but not of the magnitude we would like. As well, increased coursework in the domain is often not strongly related to pretest–posttest gains. Faculty reactions and explanations for such results are provided in Mathers et al. (2018).

THE JMU ASSESSMENT DAY MODEL

Between 3,800 and 4,800 students are required to attend each Assessment Day, with incoming first-year students (excluding transfer students) tested during Fall Assessment Day and students with 45–70 credit hours tested during Spring Assessment Day. Rather than relying on volunteers or convenience samples, JMU requires all qualifying students to participate in Assessment Days. This helps us represent students who have taken different academic paths and ensures that our results are fully reflective of the JMU experience. If a student is required to participate and fails to do so, a hold is placed on their record, prohibiting modifications to their current schedule and future course registration. This policy not only demonstrates to students and other stakeholders JMU's strong commitment to quality assessment, but also ensures participation. Fortunately, attendance is high with the 5-year attendance rate on Fall and Spring Assessment Days being 94% and 90%, respectively.

The current Assessment Day structure includes three 2-hour testing sessions, with the sessions each separated by about an hour. During each session, one third of the required students (1,200–1,500 students) are tested. To accommodate this number of students in a single session, about 25 different rooms are used, with each room seating between 30 and 170 students. Almost all rooms are located within a single building, which allows our team to be on hand to address any issues. Testing rooms are reserved more than a year in advance and include large lecture halls, small classrooms, and computer labs. To illustrate, the rooms used during Spring 2016 are listed in Table 5.

In the fall, commandeering almost an entire building is not an issue because Assessment Day takes place the Friday before classes begin. Spring Assessment Day, however, takes place on a Tuesday in mid-February; to avoid scheduling conflicts, all classes are cancelled until 4:00pm. This not only frees space on campus for university-wide assessment, but also allows students who are not required to participate in Assessment Day to participate in academic program assessment.

As many as 25 different assessments are administered on Assessment Day; each student completes no more than four assessments during their 2-hour testing session (see Table 6). Thus, large random samples of students

Table 5. Master Plan for Spring Assessment Day, 2016

Room Number	Room Size[a]	Room Size x 3[b]	Test Config.	No. of Proctors[c]	Use[d]	Test1[e]	Test2	Test3	Test4	Time1[f]	Time2	Time3	Time4	Total Time[g]	ID Range[h] Session A	Session B	Session C
0000	101	303	1	3	CBT	SDA-7	OCP2	SD-3	SOS-2	30	30	30	5	120	000–022	344–365	679–698
0201	30	90	2	2	CBT	INFOCORE	OCP2	SD-3	SOS-2	30	30	30	5	120	023–027	366–371	699–703
248	31	93	2	2	CBT	INFOCORE	OCP2	SD-3	SOS-2	30	30	30	5	120	028–033	372–378	704–711
2037	33	99	3	2	CBT	STPA2	SDA-7	SD-3	SOS-2	45	30	30	5	135	034–041	379–383	712–719
250	34	102	4	2	CBT	STPA2	OCP2	SD-3	SOS-2	45	30	30	5	135	042–050	384–391	720–726
336	30	90	5	2	CBT	ER-WRA	OCP2	MFLS	SOS-2	60	30	15	5	135	051–056	392–395	727–732
343	30	90	6	2	CBT	ER-WRA	ERRT	SD-3	SOS-2	60	10	30	5	130	057–064	396–400	733–740
350	48	144	7	2	CBT	ISNW-A1	ERRT	SD-3	SOS-2	50	10	30	5	120	065–074	401–408	741–748
2204	48	144	7	2	CBT	ISNW-A1	ERRT	SD-3	SOS-2	50	10	30	5	120	075–084	409–416	749–756
2203	40	120	8	2	P&P	ERIT-XA	SDA-7	SD-1	SOS-2	50	30	15	5	125	085–095	417–423	757–765
136	90	270	8	2	P&P	ERIT-XA	SDA-7	SD-1	SOS-2	50	30	15	5	125	096–111	424–446	766–783
148	48	144	8	2	P&P	ERIT-XA	SDA-7	SD-1	SOS-2	50	30	15	5	125	112–124	447–455	784–794
1302	160	480	9	5	P&P	NW9X	AMEX3	SD-1	SOS-2	30	40	15	5	115	125–155	456–493	795–822
2301	160	480	9	6	P&P	NW9X	AMEX3	SD-1	SOS-2	30	40	15	5	115	156–191	494–527	823–857
159	170	510	10	5	P&P	NW9	AHQ	SD-1	SOS-2	60	12	15	5	117	192–229	528–566	858–897
1301	126	378	11	4	P&P	ISNW-A1	AHQ2	SD-1	SOS-2	50	27	15	5	122	230–263	567–605	898–929
1204	50	150	11	2	P&P	ISNW-A1	AHQ2	SD-1	SOS-2	50	27	15	5	122	264–277	606–613	930–943
1209	55	165	12	2	P&P	CAT	MFLS	SD-1	SOS-2	60	15	15	5	120	278–288	614–626	944–956
1210	60	180	13	2	P&P	GLEX2	KWH8	SD-1	SOS-2	30	40	15	5	115	289–305	627–641	957–972
2209	55	165	13	2	P&P	GLEX2	KWH8	SD-1	SOS-2	30	40	15	5	115	306–320	642–655	973–985
236	80	240	13	3	P&P	GLEX2	KWH8	SD-1	SOS-2	30	40	15	5	115	321–343	656–678	986–999

[a] The number of students the room can accommodate at one time.
[b] The number of students the room can accommodate across the three testing sessions.
[c] The number of proctors needed in the room.
[d] Whether the room is used for computer-based testing (CBT) or paper-and-pencil testing (P&P).
[e] The names of the measures to be administered first in each room.
[f] The testing times for the first measure.
[g] The sum of each of the testing times for the four measures plus 25 minutes, which is the time needed to orient students to the testing session and collect or disseminate testing materials. In some rooms, Total Time exceeds 120 minutes; test configurations are allowed to exceed 120 minutes if we know, based on previous experience, that the session is not likely to take more than 120 minutes. Regardless of the projected total testing time, in practice we instruct proctors not to let testing sessions last more than 120 minutes.
[h] The range of the last three digits of student IDs assigned to each room and session (e.g., if the last three digits of a student's ID were 405, the student would report to room number 350 for Session B).

Table 6. Assessments Administered in Spring 2016 and Sample Size

Acronym	N	Name	Content Area
AHQ	510	Arts & Humanities Questionnaire	Arts & humanities
AHQ2	528	Arts & Humanities Questionnaire, version 2	Arts & humanities
AMEX3	960	The American Experience, version 3	American history & political science
CAT	165	Critical-thinking Assessment Test	Critical thinking
ERIT-XA	534	Ethical Reasoning Identification Test, version XA	Ethical reasoning
ERRT	234	Ethical Reasoning Recall Test	Ethical reasoning
ER-WRA	180	Ethical Reasoning, Writing, version A	Ethical reasoning
GLEX2	585	The Global Experience, version 2	Global history & issues
INFOCORE	183	Information Literacy Core	Information literacy
ISNW-A1	528	Institute for Stewardship of the Natural World, version A1	Environmental stewardship
KWH8	585	Knowledge of Wellness and Health, version 8	Wellness & health
MFLS	165	Meaningful Life Survey	Purpose & meaning in life
NW9	510	Natural World, version 9	Quantitative & scientific reasoning
NW9X	960	Natural World Short Form, version 9	Quantitative & scientific reasoning
OCP2	486	Oral Communications Pretest, version 2	Oral communication
SD-1	3282	Student Development, version 1	Student development
SD-3	1065	Student Development, version 3	Student development
SDA-7	534	Sociocultural Domain Assessment, version 7	Sociocultural understanding
SOS-2	4437	Student Opinion Survey, version 2	Examinee motivation
STPA2	201	Sociocultural Thought Process Assessment, version 2	Sociocultural reasoning

Note. Seventy percent of the assessments listed here are direct measures of student learning (as opposed to self-report measures of learning or self-report measures of attitudes, feelings, or behaviors). With the exception of the CAT, the direct measures listed here were created by faculty at the university.

complete each assessment, but no student completes all assessments. Assessing every student on all outcomes is not necessary because the data are not used for individual assessment purposes. The vast majority of assessments are used for program assessment purposes and are direct measures of student learning.[3] New and revised assessments are also piloted and evaluated for future use. This is particularly important because many of our assessments are developed by our own faculty and staff to maximize the alignment between program outcomes and instruments. Because the responsibility for the psychometric evaluation of these assessments falls on us, a small proportion of Assessment Day data is devoted to this purpose.

Data are also collected for the psychometric evaluation of instruments developed outside of JMU. Importantly, validity studies are conducted to ensure instruments are appropriate for use with our student population and for the purposes of program assessment. Examples of how Assessment Day data have been used in psychometric evaluations are provided by Brown, Finney, and France (2011), Cameron, Wise, and Lottridge (2007), Kopp, Zinn, Finney, and Jurich (2011), France, Finney, and Swerdzewski (2010), Johnston and Finney (2010), Smiley and Anderson (2011), and Taylor and Pastor (2007).

Planning for Assessment Day

Planning for each Assessment Day begins months in advance with the creation of a spreadsheet known as the master plan that details which assessments and student identification numbers are assigned to the various rooms and sessions (see Table 5). In the section below, we describe how and when these decisions are made, and from whom we gather the necessary information.

One of the first tasks involved in planning for a Fall Assessment Day is deciding which assessments to administer.[4] Four months prior to Fall Assessment Day, assessment coordinators for general education programs and university-wide initiatives are asked to provide information about the measure(s) that their university-wide program wishes to administer. We ask for the length of time it will take to complete the instrument(s), whether computer-based or paper-and-pencil administration is preferred, and the desired sample size. We then create test configurations based on this information (i.e., sets of three to four measures that can be given together and require slightly less than a total of 2 hours to complete).

Once the configurations are determined, we assign configurations to each testing room. In each room the same test configuration is used across each of the three testing sessions for two reasons. First, because proctors remain in the same room across sessions, keeping the test configuration consistent helps to avoid proctor confusion. Second, in paper-and-pencil testing rooms students provide their responses on Scantrons (i.e., optical answer sheets); as such, the paper copies of the tests remain unmarked and the same paper copies of tests can be reused across sessions. This helps keep the number of printed test copies to a minimum, which helps reduce costs and keep Assessment Day environmentally friendly.

The final step is to assign students to rooms and sessions based on the last three digits of their student identification numbers, as shown in the last three columns of Table 5.[5] Because the last several digits of identification numbers are used to assign students to rooms, the sample of students assigned to each room, and subsequently to each test, is random.

The above description characterizes the planning involved for Fall Assessment Days. When developing a master plan for Spring Assessment Days, we use the plan previously configured for

---

[3] A direct measure of student learning tests a student's knowledge and skills. For example, rather than asking students to self-report whether they are skilled in information literacy, we use a knowledge test to evaluate whether students are skilled in information literacy.

---

[4] Every general education program and university-wide initiative is assessed on every Assessment Day. If there is any concern about whether a program should be assessed, guidance is obtained from the university's Assessment Advisory Council, which is a team of administrators, faculty, and staff whose purpose is to provide guidance on these very issues.

the same cohort of students for their Fall Assessment Day, modifying as necessary. This helps ensure Spring Assessment Day students are assigned to complete the same measures as when they were incoming first-year students.

Human Resources

Substantial human resources are needed to orchestrate each Assessment Day. In this section, we describe two essential groups: the Assessment Day team that works year-round on the planning, coordination, and execution of each Assessment Day; and the Assessment Day proctors.

The Assessment Day Team

The Assessment Day team is a subgroup of the Center for Assessment and Research Studies (CARS), which is the unit on campus responsible for providing guidance regarding the assessment of student learning and developmental outcomes.[6] The Assessment Day team is responsible for planning and coordinating both Assessment Days, as well as for the associated data management that occurs afterward. It consists of a faculty lead, three graduate assistants (GAs), and an administrative assistant. Additionally, the team relies heavily on the CARS's information security analyst, fiscal technician, and three undergraduate work–study students to assist in tasks crucial to a successful Assessment Day (e.g., storing data securely, processing paperwork for paying proctors, packing and double-checking materials).

No member of the Assessment Day team devotes their entire work week year-round to Assessment Day. The current faculty lead of Assessment Day devotes 8–10 hours per week, on average. During the fall and spring semesters one GA on the team has 20 hours per week assigned to Assessment Day, and the remaining two GAs have 10 hours per week. The work–study students assist during the fall and spring semesters, with each of the students spending about 8 hours per week on Assessment Day tasks during the busiest times of the year.

The work associated with Assessment Day is not constant throughout the year; it is heaviest the 2 months before and after each Assessment Day. Each member of the team has different responsibilities prior to, during, and after Assessment Day, which are described below. The tasks typically completed by the work–study students during these times are also provided.

Prior to Assessment Day

Many of the tasks completed prior to Assessment Day were detailed above in the planning section. Examples include soliciting and organizing test requests, compiling test instructions, communicating with students and constituents on campus, printing proctor materials, and packing bins. These tasks are split among the GAs, followed by a rigorous round of quality checks, some of which are completed by the faculty lead and the work–study students. Prior to Assessment Day, the administrative assistant reserves testing rooms, hires proctors, and coordinates meal services. The faculty lead is primarily responsible for coordinating work among the team members and ensuring that work is completed by the prespecified deadlines.

During Assessment Day

During Assessment Day the administrative assistant oversees the completion of paperwork for hiring

---

[5] Specifically, we begin by acquiring the list of student identification numbers for all incoming first-year students and sort this list by the last three digits of the identification number. Starting with a value of 000, we assign three-digit values to rooms and sessions, starting with the first room and Session A. Once the number of students reaches the room size, we progress to the next room. After we have progressed through all rooms for Session A in this manner, we repeat the process for Session B and then Session C. Starting in Fall 2018 we began assigning students based on the last four digits of their identification number (instead of three digits) to accommodate increases in the size of the student body.

[6] At our university, the assessment office (CARS) and the Office of Institutional Research are separate and the latter does not assist with Assessment Days. In many universities, assessment falls under the purview of an institutional research office or a larger strategic planning office. How feasible it is to implement the Assessment Day model in these different organizational configurations depends on the number of staff, size of the student body, and the scope of assessment (e.g., number of assessments, number of Assessment Days).

proctors, coordinates delivery of meals, and answers the phone in the room that serves as headquarters. The faculty lead welcomes the proctors and answers questions. Once proctors proceed to their designated rooms and students begin to arrive, two of the GAs act as runners who move throughout the testing rooms to help proctors set up. The third GA and the faculty lead remain in headquarters to respond to any other needs and to monitor the CARS email account for student questions. The CARS information security analyst is also present in headquarters to assist with technology issues. After the final testing session, the team collects materials, packs up the headquarters room, checks all testing rooms for any forgotten materials, and ensures rooms are left the way they were found.

After Assessment Day

After Assessment Day the GAs oversee the work–study students in the unpacking of all materials (e.g., Scantrons, tests, pencils, folders, bins, binders) and their inventory. The work–study students also check technology, such as Chromebooks (i.e., tablet-like laptops), to ensure that everything is in working order. In sum, the work–study students help us ensure that all materials are accounted for and ready for future use.

Scanning and downloading of data is completed within a week of Assessment Day, thereby allowing the team to track attendance. Students who failed to attend (either for legitimate reasons or out of delinquency) have a hold placed on their record and are contacted via email about make-up sessions. There are typically two to six make-up sessions, each accommodating about 100 students, scheduled in the evenings several weeks after Assessment Day. The GAs plan and proctor the make-up sessions, and the administrative assistant removes holds for students who attend.

The management of all data also occurs within a month after Assessment Day and includes data scanning, downloading, cleaning, scoring, and formatting. Using the student identification numbers supplied by the student on each assessment, the data are also merged with other information needed for program assessment purposes; for instance, assessment scores for each student are merged with relevant course information. All GAs aid in data management and subsequent quality checks. Each program's assigned assessment liaison (with assistance from their own GAs) completes the analyses and report writing for each assessment within 3 months of testing.[7] Results are reported to the program faculty and staff, who may choose to disseminate the results more widely. Although it varies across programs, faculty and staff often meet to discuss the results and consider potential changes to their program. They are encouraged to use a learning improvement model, where assessment results obtained after program changes have been made are used to determine if the changes were effective (Fulcher, Good, Coleman, & Smith, 2014).

Assessment Day Proctors

Proctors are an important human resource that we greatly rely on. Although the number of proctors varies, our goal is to have one proctor for every 30 students with no fewer than two proctors in a room, which results in about 55 to 75 proctors. Proctor recruitment begins 2 months before Assessment Day when the team's administrative assistant emails a job announcement and online application form to a list of potential proctors (including JMU graduate students, staff, and people who have previously served as proctors). We have many people in the local community who regularly proctor, many of whom are retired educators. From this referral-based network, completed applications are selected on a first-come, first-served basis. The application is closed once we have enough proctors, which typically occurs within 3 weeks. Proctors are paid a small stipend and are provided breakfast and lunch on Assessment Days.

Because there are at least two proctors per room, it is important that proctors within a room act as a team. To facilitate cooperation, one proctor is assigned to be lead proctor; he or she acts as the spokesperson to the students, directs the testing session, and delegates tasks among other proctors. Both lead and non-lead proctors are responsible for a variety of other tasks. For instance, proctors are responsible for preparing the room for each session and

[7]*Care is taken in reporting so that the results can only be used to evaluate programs, not individual students or faculty members.*

maintaining order (e.g., minimizing noise, disruptions, and inappropriate behaviors). Proctors also convey the importance of the assessments and create an environment that allows and encourages students to perform to the best of their ability. Thus, proctors have an important role in ensuring the quality of the data: they motivate students, ensure tests have been completed correctly, and report any noteworthy issues that could impact the results. How proctors are trained to accomplish these tasks is briefly described below and in more detail by Lau, Swerdzewski, Jones, Anderson, and Markle (2009).

Changes Made to Assessment Day

JMU's Assessment Day model has evolved over time. Many changes have been made in response to increases in the size of the student body, developments in testing technology, and issues encountered after implementing an Assessment Day. Because our model has been in place for more than 30 years, it is impractical to describe all of the changes that have been made. We focus here on large changes that have improved the quality of data, saved money, improved efficiency, or reduced the environmental impact of Assessment Day.

Number of Testing Sessions

Perhaps the most significant change made in recent years is the transition from two 3-hour testing sessions to three 2-hour testing sessions. This change allowed the number of students tested to be distributed over three sessions instead of two, thereby requiring fewer rooms, proctors, and testing materials. For instance, when the Spring 2015 administration, which used the three 2-hour testing session structure, was compared to the Spring 2014 administration, which used the two 3-hour structure, substantial decreases were noted in the number of proctors (↓38%), Scantrons (↓45%), and copies of assessments (↓56%). Not only did this change reduce the amount of time required by any one student for testing, but it also greatly reduced costs as well as the environmental impact of Assessment Day.

Assessment Day Video

Beginning in 2014, we started to show students a 5-minute video at the beginning of each testing session; in this video the president of the university, general education faculty, and student actors explain the purpose of Assessment Day. The purpose of showing the video is two-fold: to increase student motivation and to standardize how information is communicated. By informing students how the data collected on Assessment Day are used to improve student learning on campus we hope to convey how completing the assessments to the best of their ability directly affects the quality of education at JMU as well as its reputation. Readers interested in viewing the video can find the link at JMU (2018, top of page).

Proctor Selection and Training

A few years ago, we made modifications to the way we recruit and hire proctors. We converted our proctor-hiring methods from an informal email process to a formal online application. Under the new hiring method, proctors complete an online application that allows us to collect necessary information before Assessment Day. The online application has also allowed us to ensure that our proctors are comfortable with technology. As we move to testing that is more computer-based, we need proctors who can navigate various types of technology with ease. By having proctors apply through an online form, we create a preliminary screening process for this skill. Additionally, we have started recruiting JMU graduate students to serve as proctors, which provides many benefits. Graduate students are generally familiar with and comfortable navigating JMU's campus and classroom technology and they usually have less hiring paperwork to process because they are often already JMU employees (e.g., GAs). The quality of proctors is somewhat controlled by our detailed job description and online application process. The Assessment Day team also observes proctors during Assessment Day and does not rehire proctors who perform poorly.

Another notable change we have made to Assessment Day is to the timing and format of proctor training. At one time proctors were trained the morning of Assessment Day; however, the training session added an hour to an already long day and was often rushed. There was a lot of information packed into a quick presentation, leaving little time for proctors to reflect on the material and ask questions before being ushered into their rooms. To address these challenges, we moved the training online, which allows us to track which proctors have completed training and allows proctors

to complete the training in their own space and time during the 2 weeks prior to Assessment Day.

Ongoing Challenges

Student Motivation

The primary purpose of Assessment Day is to collect meaningful information about what students know, think, and can do. Our ability to make valid inferences from students' scores relies on the quality of the data we collect. Unfortunately, the quality of the data is undermined when students are not motivated. Although we attempt to convey the purpose and importance of Assessment Day to students, the assessments are still low-stakes for students, and, as in any low-stakes assessment context, examinee motivation can suffer.

Concerns about student motivation are mitigated somewhat by data indicating the majority of students think the assessments are important and try their best (e.g., see Sundre & Wise, 2003, Table 2). This is particularly true of incoming first-year students. However, because these findings do not characterize all students, we are continuously looking for ways to improve motivation. One strategy we use is to train our proctors to use motivational strategies as part of their role. We began intentionally training proctors in 2007 to use motivational strategies (e.g., conveying the importance of the test, being supportive yet firm, etc.) and found that students' self-reported effort on the assessments was higher and less variable on Assessment Days that took place after this training was implemented (Lau et al., 2009).

We have also studied the effects of providing different instructions to students (Finney, Sundre, Swain, & Williams, 2016). During this study students were randomly assigned to one of three sets of instructions: In Condition 1 we told students that their scores would be aggregated and used for institutional decision-making, in Condition 2 we expanded on Condition 1 by telling students they would be able to receive their individual scores, and in Condition 3 we added to Conditions 1 and 2 by informing students that their individual scores would also be shared with faculty. Test performance from pretest to posttest along with test-taking motivation measures were not affected by the kind of instructions the student received.

We have also piloted different assessment designs, such as a planned missingness design, to investigate whether giving students a portion of the assessment rather than whole assessment can improve motivation and performance (Swain, 2015). Although the effects were small, students completing only a portion of the assessment (about 33 items) performed better than students completing the whole assessment (66 items). In addition, their motivation was more favorable, but not significantly so.

As a final example, the use of electronic pop-up messages targeted at students displaying rapid responding behavior on computer-based tests has been investigated (Ong, Pastor, & Yang, 2018; Wise, Bhola, & Yang, 2006), with mixed results regarding the effectiveness of this intervention. In addition to changes aimed at improving student motivation, we are continuously researching different ways to measure motivation (e.g., self-report, item response time; Wise & Kong, 2005), assess its impact on the inferences we make (e.g., Finney et al., 2016), and accommodate the issue during our analyses (e.g., Foelber, 2017; Sundre & Wise, 2003).

Efficiency

Another important challenge we continue to face is the issue of efficiency. We have turned to electronic data collection as a primary way of reducing both our costs and our environmental impact. We consistently prioritize the use of on-campus computer labs to reduce both the number of paper tests and Scantrons needed. Furthermore, we recently incorporated the use of Chromebooks, which are tablet-like laptops, in rooms that were formerly used for paper-and-pencil testing. This allows us to assess around 200 students outside of a computer lab but still without resorting to Scantrons. We have also experimented with having students respond via handheld survey response tools on their smartphones (Sauder, Foelber, Jacovidis, & Pastor, 2016).

With the emphasis on electronic data collection, the challenges we currently face are mostly physical limitations (e.g., number of available computer labs). A similar challenge is the lack of alternative technology for assessing students outside of the computer labs. The Chromebooks continue to be valuable in this regard, but we are limited by the number of Chromebooks we can purchase. Our experiments with handheld responding devices have

been challenged by considerations of cost and ease of use. Our pilot of student-owned smartphones had the advantage of being free for us, but brought its own challenges in terms of test security and student attention. Yet, we are optimistic about the future of technology in Assessment Day to increase our efficiency.

CONCLUSION

While the details can be dense, we hope they convey the thought and intentionality involved in our Assessment Day model. It is our hope that this information benefits institutions wanting to adopt an Assessment Day model for university-wide assessment. For institutions where our model may not be feasible or even desirable to implement on a large scale, aspects of our model can be adopted for assessment on a much smaller scale, even for the assessment of a single program. For institutions with Assessment Days already in place, we hope our description provides ideas for different ways to implement the model and alternative solutions for addressing its challenges. In sum, our intention is to share lessons learned and encourage readers to consider how our model might be adapted for their own purposes.

Although we featured our Assessment Day model, we are open and supportive to any design that facilitates the collection of quality data. In addition, we encourage any institution with a quality process to share its approach and lessons learned with others. Let's share quality practices to better answer the calls for accountability and support legitimate learning improvement efforts.

REFERENCES

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to represent the dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement*, 71, 170–185. doi:10.1177/0013164410387378

Cameron, L., Wise, S. L., & Lottridge, S. M. (2007). The development and validation of the information literacy test. *College & Research Libraries*, 68, 229–237. doi:10.5860/crl.68.3.229

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, 21(1), 60–87. doi:10.1080/10627197.2015.1127753

Fisher, R., Smith, K., Finney, S., & Pinder, K. (2014). The importance of implementation fidelity data for evaluating program effectiveness. *About Campus*, 19(5), 28–32.

Foelber, K. J. (2017). *Using multiple imputation to mitigate the effects of low examinee motivation on estimates of student learning.* Doctoral Dissertation. James Madison University, Harrisonburg, VA.

France, M., Finney, S. J., & Swerdzewski, P. (2010). Students' group and member attachment to their university: A construct validity study of the University Attachment Scale. *Educational & Psychological Measurement*, 70, 440–458. doi:10.1177/0013164409344510

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. Occasional Paper# 23. National Institute for Learning Outcomes Assessment, Champaign, IL.

Gerstner, J. J., & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research & Practice in Assessment*, 8, 15–28.

Grays, M. & Sundre, D. L. (2012, November). *Lessons learned from 25 years of Assessment Days*. Presented at the Virginia Assessment Group Annual Conference, Richmond, VA.

Harris, H., & Horst, S. J. (2016). A brief guide to decision at each step of the propensity score matching process. *Practical Assessment, Research, & Evaluation*, 21(4). Available at http://pareonline.net/getvn.asp?v=21&n=4

Hathcoat, J. D., Sundre, D. L., & Johnston, M. M. (2015). Assessing college students' quantitative and scientific reasoning: The James Madison University story. *Numeracy*, 8(1), Article 2. doi:http://dx.doi.org/10.5038/1936-4660.8.1.2

James Madison University (JMU). (2018). About Assessment Day. Harrisonburg, VA. Available at https://www.jmu.edu/assessment/Students/aboutAday.shtml

Johnston, M. M. & Finney, S. J. (2010). Measuring basic needs satisfaction: Evaluating previous research and conducting new psychometric evaluations of the Basic Needs Satisfaction in General Scale. *Contemporary Educational Psychology*, 35, 280–296. doi:10.1016/j.cedpsych.2010.04.003

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the Academic Entitlement Questionnaire. *Measurement and Evaluation in Counseling and Development*, 44, 105–129. doi:10.1177/0748175611400292

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee efforts on general education program assessments. *Journal of General Education*, 58, 196–217. doi:10.1353/jge.0.0045

Mathers, C. E., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*, 1–17. doi:10.1080/02602938.2018.1443202

Ong, T. Q., Pastor, D. A., Yang, S-T. (2018, April). The effects of administering alerts at fixed points during a low-stakes test. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Pieper, S. L., Fulcher, K. H., Sundre, D. L., & Erwin, T. D. (2008). "What do I do with the data now?": Analyzing assessment information for accountability and improvement. *Research & Practice in Assessment*, 2, 1–8.

Sauder, D. C., Foelber, K. J., Jacovidis, J. N., & Pastor, D. A. (2016, Summer). Utilizing tech-

nology in data collection. AAHLE Intersection, 7–9.

Smiley, W. F., & Anderson, R. D. (2011). Measuring students' cognitive engagement on assessment tests: A confirmatory factor analysis of the short form of the Cognitive Engagement Scale. *Research & Practice in Assessment*, 6, 17–28.

Sundre, D. L., & Wise, S. L. (2003, April). *"Motivation filtering": An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the annual meeting of the National Council on Measurement in Education Annual, Chicago, IL.

Swain, M. (2015). *The effects of a planned missingness design on examinee motivation and psychometric quality* (Doctoral dissertation). James Madison University, Harrisonburg, VA.

Swain, M. S., Finney, S. J., & Gerstner, J. J. (2013). A practical approach to assessing implementation fidelity. *Assessment Update*, 25(1), 5–7.

Swing, R. L. (2001). Dedicated Assessment Days: Mobilizing a campus's efforts. *Assessment Update*, 13(6), 13–15.

Taylor, M. A., & Pastor, D. A. (2007). A confirmatory factor analysis of the Student Adaptation to College Questionnaire. *Educational & Psychological Measurement*, 67, 1002–1018. doi:10.1177/0013164406299125

Wise, S. L., Bhola, D. S., & Yang, S. (2006). Taking the time to improve the validity of low stakes tests: The effort monitoring CBT. *Educational Measurement: Issues and Practice*, 25(2), 21–30. doi:10.1111/j.1745-3992.2006.00054.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183. doi:10.1207/s15324818ame1802_2