University of Delaware Questions and Feedback on the DDOE Scorecards

Program Scorecard Feedback

1. There should be an Elementary report that is not also tied to a concentration.

2. Candidate Performance: Average SAT Score of Incoming Class.

We provided the average SAT score and the number of candidates upon which the score was based. However, the Health and Physical Education scorecard did show the SAT score when the number of candidates was less than 10; this score should not appear. In addition, the Elementary/Middle School English and Elementary/Middle School Science scorecards did not show the SAT score when the number of candidates was more than 10. All Elementary/Middle School reports should show an average SAT of 1762.

The average SAT score is inconsistent with the Council for the Accreditation of Educator Preparation (CAEP) requirements relative to SAT information. Why is there a lack of consistency between your metric and CAEP's metric?

3. Placement rate in Delaware. More than 10 ARTC candidates worked in Delaware after graduating in 2013. The ARTC scorecard should have a placement rate for this metric.

4. Retention. ARTC candidates are full-time teachers in Delaware classrooms while they earn their certification, so it is not clear when their time in the classroom begins (as a candidate or after they earn certification). For example, would the retention rate beyond year one be 100% for ARTC candidates because they are in the classroom for at least two years?

5. Graduate Performance. The Agriculture, Elementary/Middle School Science, Elementary/Middle School English, and Foreign Language scorecards have Student Improvement Ratings, but they do not have Observation Scores due to "fewer than ten educators in sample." Please explain how there could be more than ten educators for the Student Improvement Ratings and less than ten educators for the Observation Scores when both sets of data come from the same DPAS- II evaluations. In addition, these four programs do have Overall Performance Evaluation Ratings from DPAS-II.

6. About This Program.

The contact, department, and department chair information is incorrect for almost all of the UD programs. Specifically, the information for one program appears in a report for a different program (e.g., the Health and Physical Education report has the Foreign Language Education information). Please review the information that we originally sent.

We are not sure why the CAEP and SPA categories say “N/A” when we were are NCATE-accredited and are nationally recognized by all national specialty program associations (SPAs). It seems as if “NCATE” and the specific SPAs (e.g., NCTM) should be listed in these categories, respectively.

We provided the number of full-time tenure-track faculty and the number of full-time continuing (non-tenure) track faculty for the undergraduate and graduate programs in July. These two tracks are full-time, and their numbers should be summed for the Tenure/Full Time category. Please sum these two sets of data, which will result in percentages that add up to 100% (the percentages are currently less than 100%).

Please update the ARTC faculty numbers to Total Faculty: 2, Tenure/Full Time: 2, and Adjunct: 4.

The 3-year trend of spring graduates on the ARTC scorecard is incorrect because the numbers are too high.

We provided the average SAT/ACT scores, average GPA scores, and the number of candidates upon which the scores were based. However, the Health and Physical Education and Foreign Language Education scorecards showed the average GPA when the number of candidates was less than 10; their average GPA should not be listed.

The GPA on the ARTC scorecard should be higher than 3.0 because the minimum GPA admission requirement for ARTC is 3.0.

The Total Enrollment category should be deleted under the “2014-2015 Graduates” heading because this category shows the total enrollment in the program for Fall 2014, not the number of graduates; the category does not match the heading.

The “Graduates” heading on the ARTC scorecard does not list the year(s) for the data.

We provided data on the 2013-2014 graduates, so the heading “2014-2015 Graduates” should read “2013-2014 Graduates.”

The MPCP program field experience hours should be, “Two years full-time.”

Please change the faculty numbers to the numbers listed below.

Program	Total	Tenure/Full Time	Adjunct
English	10	10	0
Foreign Languages	10	10	0
Secondary Mathematics	13	13	0
Music	7	7	0
Social Studies	9	9	0

Agriculture	8	8	0
Health and Physical Education	12	12	0
Blended Early Childhood Education	14	14	0
Elementary and Middle School English	39	36	3
Elementary and Middle School Math	39	35	4
Elementary and Middle School Science	38	35	3
Elementary and Middle School Social Studies	38	35	3
Elementary and Special Education	44	41	3
Exceptional Children and Youth (MPCP)	14	11	3

Technical Specifications Questions

1. How were the following selected? What research supports their selection?
- Six domains
- Minimum standards and state targets
- Weights for each domain and metric

2. To follow up on Q1, we are particularly concerned with the Placement and Retention domains. These are not under the control of any teacher preparation program. We have no influence on the hiring or retention of our candidates; these are decisions totally in the hands of districts and schools. Together, these domains represent about 35% of the total score (30/85).

3. As a second follow-up to Q1, we are concerned about the Graduate Performance metrics. We endorse the inclusion of this domain, but question the four metrics (i.e., why include the overall rating when it is represented in the first three, and the relative weights assigned to each).

4. As a third follow-up to Q1, we are concerned about the logic for the number of years of data considered for the two Placement metrics. Only one year of data is used for the first (Placement Rate in Delaware) and five years for the second (Placement Rate in Delaware High-Needs Schools). Since both metrics depend on the same data set, the second is a subpart of the first, why not report data on both metrics for the same time period?

5. We are particularly confused about the logic/rationale that supports the four metrics for the Graduate Performance domain. Explain the decision to not consistently use the 10th and 90th percentile in reporting Delaware's historical performance.

6. For the input metrics, Percent of Candidate Class, Non-White and Average SAT Score of Incoming Class, the 2015 Technical Specifications considers all of a preparation program's incoming students. And for most of the output metrics, only candidates teaching in Delaware are considered. This makes sense to us. However, for one of the output metrics, Average Score on the General Knowledge Exam, the text in the 2015 Technical Specifications gives conflicting information (see p. 6). In particular, the Metric Description specifies only "program graduates who have worked within public education in Delaware"

while the Universe description indicates “educators graduating from an identified educator preparation program between…” Please clarify.

7. We are also confused by the text related to the Overall Performance Evaluation Ratings (see p. 17). The text in the Example Calculation suggests that this metric relies only on the Student Improvement Component of DPAS-II instead of the overall rating. Please clarify.

8. We found multiple errors in the UD reports. What procedures are in place to verify the accuracy of the data included in these reports? What procedures are available for teacher preparation programs to review the reports for accuracy before they are released publicly?

9. The reports include no information on sample sizes or missing data rates for any metrics, making impossible to determine the statistical precision of the metrics. Can the reports be revised to show analytic sample sizes for each metric?

10. The lower limit of 10 teachers for reporting may be large enough to ensure confidentiality for individual teachers, but it is likely not large enough to ensure adequate statistical precision for any given metric. What is the rationale behind the lower limit of 10 teachers?

11. What is the formula for the “variance weighing” of VAM scores under the Student Growth Outcomes metric. Can you cite references from published literature?

12. Given that data on some subset of metrics may be missing for any given program, the metrics and ratings produced for any one program are not comparable to any other program for which a different subset of metrics are available. For example, programs with insufficient VAM data for Student Growth Outcomes may have metrics and ratings that are biased upwards. Have you performed any non-response bias analyses to evaluate whether this is happening or not?

13. Why is your graduate performance metric not consistent with what CAEP is requiring of us?

14. We did not receive the document providing support for the domains, metrics, and benchmarks in time to provide a response. We know, based on research and the Delaware Department of Education's own comments, for example, that the research does NOT support placing novice teachers in high-needs schools.
Yet, it is included as a metric.

Issues

1. Selection of proposed scorecard domains – Six Domains
a. Are these the best indicators to measure the effectiveness of our teacher education programs? It would be helpful for some evidence to support their selection.
b. Fully supports two: Recruitment and Candidate Performance.

c. Two are more complex: Graduate Performance and Perceptions. Many factors besides teacher preparation programs have been shown to influence candidates' performance on these. The state proposed metrics do not adequately reflect the complexity of these performance domains. (need to say more)

d. Two are outside our control: Placement and Retention. UD has no control over who is hired and whether they are retained.

2. Standard Setting, Percentiles, and Years of Data

a. The process and rationale for setting minimum standard and state targets is not clear. For example, 1550 is widely used for SAT cut-off score for college-ready students. However, it's not clear how "1950" was set as the state target, seems fairly ambitious given distribution of SAT scores. For other domains, the rationale is even less clear. We do not disagree with the need for setting minimum and state targets, but the process needs to be transparent, tied to research/best practice, and goals set explicitly by the preparation programs and the DDoE.

b. The weights for the four metrics for the Graduate Performance domain need further clarification. We particularly question the assignment of substantial weights to the Student Improvement Component Ratings given concerns raised about the validity and reliability of the collection of non-DCAS measures. In addition, it is not clear from the text whether the Overall Performance Evaluation Rating considers all five components or just the fifth component. It's not clear what purpose is served by the fourth metric given the earlier three.

c. Along the same lines, it is not clear why the system does not consistently set the minimum level of performance as the 10th percentile (proposed ranges vary from 10-30 percentiles) and the state target as the 90th percentile (proposed ranges from 90-95 percentiles). What is the rationale for these variations?

d. The number of years of data varies substantially without much explanation. For example, in the domain of Recruitment, the percent of class non-white is based on five years of data while the average SAT is calculated on only a single year of data. Similarly, for the Placement domain, the placement rate of candidates is based on a single year, while the placement of candidates in high needs schools is based on five years of data. The latter example is particularly puzzling because the second metric is a simple subset of the first metric.

e. Text descriptions are not consistent and clear. For example, on the Candidate Performance metric – Average Score on the General Knowledge Test, the technical specifications say that the universe is "all program graduates who have worked within public education in Delaware" (Metric Description) while later text indicates that the universe is "educators graduating from an identified educator preparation program between years 2009 and 2013" (Universe Description). It is important that text specification are clear and consistent throughout.

f. It would also be helpful for the DDoE to explain how the ranges were established for Tiers 1, 2, 3, and 4.

3. Statistical Issues

a. The lower limit of 10 teachers for reporting may be large enough to ensure confidentiality for individual teachers, but it is likely not large enough to ensure adequate statistical precision for any given

metric. Given the high stakes associated with these reports, it is imperative that sample sizes be adequate to ensure that comparisons between programs are not unreliable. For example, if a program has 20 graduates teaching in DE, and 10 of these are non-white, the 95% confidence interval for the proportion non-white is 50% ± 22%.

b. Given that data on some subset of metrics may be missing for any given program, the metrics and ratings produced for any one program are not comparable to any other program for which a different subset of metrics are available. Simply allocating the points from a metric having missing data to the other metrics in the domain may produce significant bias in the results. The current method for handling missing data imposes the assumption that performance on the metric with missing data would be similar to performance on the other available metrics in that domain. This assumption is very likely to be violated.

c. The "variance weighting" mentioned in the technical report is not sufficiently described (i.e., no formulas are presented), nor are citations provided. How is this similar to or different from precision weighting techniques used in VAM research?

4. 2015 Educator Preparation Program Reports

Compiling and analyzing the necessary data for these reports is a tremendous task. We found numerous errors in the reports that raise questions about the DDoE's work. For example:

a. Text in the About This Scorecard indicates that five years of program data were considered. This is not accurate, the range of years of data varies from 1-5.

b. We calculated the points earned for each metric and the tier ratings using the data reported. In general, our calculations came fairly close to those reported for the former once we considered rounding errors. Our calculations revealed larger gaps in the overall tier ratings. That said, it would be helpful to report program percentages with at least one or two decimal places so that the precision in numbers were more comparable.

c. Years of data for Graduate Performance: Student Growth Outcomes metrics is not reported. Only four years of data were available. This should be noted.

d. In About this Program, the data reported should generally match or align with the data reported in the previous pages. All of the data (e.g., admissions, accreditation, faculty, field experience, number of graduates) should be reported for the five years included in the report. More clarity is needed about what is being reported. For example, what do the ranges for the program SAT/ACT signify? Are these the lowest and highest, the range across a particular number of years, and so on.

e. In About this Program, number of spring graduates for three years (in graph) does not match the data reported in the table of 2014-15 graduates. It would be helpful if the graph and table matched and the two matched the data used to determine the program's performance on the various metrics.

f. In About this Program, the correct contact information is not always included. For example, the contact for UD English Education (Secondary) is not Dr. Ferretti.

g. In a specific report, UD Elementary Teacher Education/Middle School English – data are not reported for the Student Growth Outcomes or the Observation Scores because of small sample sizes. It is not clear to us how data could then be presented for the Student Improvement Component Ratings or for the Overall Performance Evaluation ratings.

h. The report includes section on Assurances of Delaware Requirements. This is not explained in the Technical Considerations text.

It is important that these reports are accurate, both in terms of the findings as well as the text. Reports like these stand or fall on the credibility of their data. It is important that sufficient time be taken to verify and accurately report all narrative text and data.

September 15, 2015

Dear Interim Secretary of Education Godowsky:

The University of Delaware applauds the Delaware Department of Education for undertaking the review of educator preparation education programs in the state. This is important and demanding work, identifying indicators of success that can be collected in valid and reliable ways across multiple programs with limited resources is challenging. Reaching consensus on specific indicators across multiple stakeholders is also difficult. We appreciate the investment of state resources in this undertaking. The resulting data will be useful to potential students and their families, to districts and schools making hiring decisions, and to the programs themselves in reviewing and improving their educator preparation programs.

In our review of the Department's work to date, we have the following comment and concerns.

1. We support the generation and publishing of educator preparation program reports. This is important and valuable work for the Delaware Department of Education to undertake. We also commend the Department for including both input and output metrics.

2. The University of Delaware objects to the inclusion of the Placement and Retention Domains. These two domains are not within the control of UD's educator preparation programs. Instead, individual districts and schools make placement decisions, and educators' retention in those placements depends on a variety of both professional and personal factors. These two domains account for about 35% of the total score. This is unacceptable to UD.

3. The University of Delaware suggests that the Department be more transparent in its work on the specific domains, metrics and relative weights, and the setting of minimum standards and state targets. It is not clear what research or best practice informed these decisions. In some cases, the decisions seem arbitrary and not anchored to any research.

4. Several aspects of the statistical analyses are problematic. Most importantly, (a) the sample sizes for most calculations are too small to allow reliable comparisons between programs, and (b) the method for handling missing data is likely to induce bias in the results, making cross-program comparisons unfair and potentially misleading.

5. Because many of the metrics used in the reports are based on data from only those teachers actually teaching in DE, the sample of teachers used to rate any given program is not reflective of the total pool of teachers who might have been recruited to teach in DE. For example, if the best teachers from a program are recruited to teach in MD or PA, then only the least effective teachers would be recruited into DE schools, and that program would receive a tier rating lower than what would truly reflect the quality of its graduates. If principals and superintendents use the reported program ""tiers"" to inform their decisions on whom to hire (e.g., preferring teachers from programs in higher tiers), then they might end up recruiting the wrong teachers by making even less of an effort to recruit the excellent teachers that DE has been losing to MD and PA. In short, the reports, as currently prepared, may have significant unintended consequences for districts' recruitment of educators.

6. These reports and the data included in them are high stakes for educator preparation programs. It is important that the data presented are accurate and complete. We found numerous errors in the data reported for our preparation programs. The Department needs to be transparent about how these reports were prepared, steps taken to verify their accuracy, and procedures for educator preparation programs to review and correct incomplete or misinformation.

7. We also have suggestions for improving the text and accompanying tables and figures in the Educator Preparation Program Reports. We volunteer to work with Department representatives to strengthen the composition of these reports.

8. The scorecard, with its domains and metrics, goes well beyond what is required in Senate Bill #51. (Please see the list below comparing what the Department elected to include in the scorecard and what is required in Senate Bill #51). Given the lack of agreement around the scorecard metrics and the insufficient time now available for teacher educators to review the research and best practice used in the creation of the scorecard metrics, we request that only those items required by Senate Bill #51 be included in the first scorecard report. We want data, for example, on our graduates' performance on the five DPAS II components.

In addition to the table comparing the scorecard domains and metrics with Senate Bill #51 requirements, I have included a list of specific technical specification problems identified by our research and evaluation experts.

I appreciate your willingness to work toward a positive solution to this problem, Steve. If I, or anyone else at UD, can be of assistance, please do not hesitate to contact me. I look forward to working with you in your new position.

Sincerely,

Carol Vukelich
Interim Dean, College of Education and Human Development University of Delaware

Comparison of the DOE Scorecard and the Senate Bill 51 Program Requirements

1. Recruitment: Percent of Candidate Class, Non-White
Reported here is the proportion of educators that are non-white amongst those that have graduated from this program in the past five years and have worked within public education in Delaware.
NOT IN SENATE BILL 51.

2. Recruitment: Average SAT Score of Incoming Class
For the 2015 scorecard, this measure represents the average cumulative SAT score (reported on a scale of 2400) for the most recent incoming class of the program. This also includes ACT scores, converted to their SAT equivalent.

SENATE BILL 51 STATEMENT
Each educator preparation program approved by the Department shall establish rigorous entry requirements as prerequisites for admission into the program. At a minimum, each program shall require applicants to: (1) Have a grade point average of at least a 3.0 on a 4 .0 scale or a grade point average in the top 50th percentile for coursework completed during the most recent two years of the applicant's general education, whether secondary or post-secondary; or (2) Demonstrate mastery of general knowledge including the ability to read write and compute by achieving a minimum score on a standardized test normed to the general college-bound population such as Praxis, SAT or ACT as approved by the Department. Each educator preparation program may waive these admissions requirement s for up to 10% of the students admitted.

3. Candidate Performance: Average Score on the General Knowledge Exam
Reported here are the average General Knowledge Exam scores for all program graduates who have worked in a Delaware public school.
NOT IN SENATE BILL 51.

4. Candidate Performance: Average Score on Performance Assessment
Reported here are the average performance assessment score(s) for all program graduates who have worked within public education in Delaware.

SENATE BILL 51 STATEMENT
Each educator preparation program approved by the Department shall establish rigorous exit requirements which shall include but not be limited to achievement of passing scores on both a content-readiness exam and a performance assessment.

5. Placement: Placement Rate in Delaware
Reported here is the rate at which program graduates begin working in public education in Delaware within one year of graduation.
NOT IN SENATE BILL 51.

6. Placement: Placement Rate in Delaware High-Needs Schools
Reported here is the rate at which program graduates begin teaching in a Delaware public school that has been state- identified as high-need.
NOT IN SENATE BILL 51.

7. Retention: Beyond Year One Retention Rate
Reported here is the rate at which program graduates continue working in public education in Delaware beyond year one.
NOT IN SENATE BILL 51.

8. Retention: Beyond Year Three Retention Rate
Reported here is the rate at which program graduates continue working in public education in Delaware beyond year three.
NOT IN SENATE BILL 51.

9. Graduate Performance: Student Growth Outcomes
Reported here are the student achievement results of program graduates teaching English, Math, Science or Social Studies.
Graduate Performance: Student Improvement Component Ratings
Reported here is the performance of program graduates' on the Student Improvement Component of their evaluation using multiple measures of student growth.
Graduate Performance: Observation Scores (Average Calculated Criteria/Component Ratings)
Reported here are the average observation scores earned by program graduates.
Graduate Performance: Overall Performance Evaluation Ratings
Reported here is the performance of program graduates' based upon their overall educator evaluation system ratings.

SENATE BILL 51 STATEMENT
Educator preparation programs shall collaborate with the Department to collect and report data on the performance and effectiveness of program graduates. At a minimum, such data shall measure performance and effectiveness of program graduates by student achievement. The effectiveness of each graduate shall be reported for a period of 5 years following graduation for each graduate who is employed as an educator in the State. Data shall be reported on an annual basis. The Department shall make such data available to the public.

10. Perceptions: Preparedness Index, Skill Survey
Reported here are the results of program graduate survey data collected within their first year of serving in Delaware's schools.
NOT IN SENATE BILL 51.

11. Perceptions: Preparedness Index, LEA 360

Reported here are the results of administrator/hiring authority survey data assessing program graduate readiness in several key performance factors in their first year.
NOT IN SENATE BILL 51.

12. About This Program – Back page. Contact Information, including phone, email and/or web site (program, department, and department chair), Accreditation (CAEP, SPA), Faculty (Total, Tenure/Full Time, Adjunct), Number of Graduates, 3-year trend, Program Description (program web site)
NOT IN SENATE BILL 51.

13. Admissions (Average SAT/ACT Score, Average GPA, Percent of Candidates Admitted Under Criteria Waiver)

SENATE BILL 51 STATEMENT
Each educator preparation program approved by the Department shall establish rigorous entry requirements as prerequisites for admission into the program. At a minimum, each program shall require applicants to: (1) Have a grade point average of at least a 3.0 on a 4 .0 scale or a grade point average in th e top 50th percentile for coursework completed during the most recent two years of the applicant's general education, whether secondary or post-secondary; or (2) Demonstrate mastery of general knowledge including the ability to read write and compute by achieving a minimum score on a standardized test normed to the general college-bound population such as Praxis, SAT or ACT as approved by the Department. Each educator preparation program may waive these admissions requirement s for up to 10% of the students admitted.

14. Field Experiences

SENATE BILL 51 STATEMENT
Each educator preparation program approved by the Department shall incorporate the following: (1) A clinical residency component supervised by high quality educators, as defined by the Department. The clinical residency shall consist of at least ten weeks of full-time student teaching.

15. 2014-2015 Graduates (Total Enrollment, Males, Females, White, Black, Hispanic, Other)
NOT IN SENATE BILL 51.