Summary Report CON 22-100

AGREEMENT FOR CONTRACTOR SERVICES BETWEEN ENTERPRISE FLORIDA, INC. AND THE ROOSEVELT GROUP, LLC

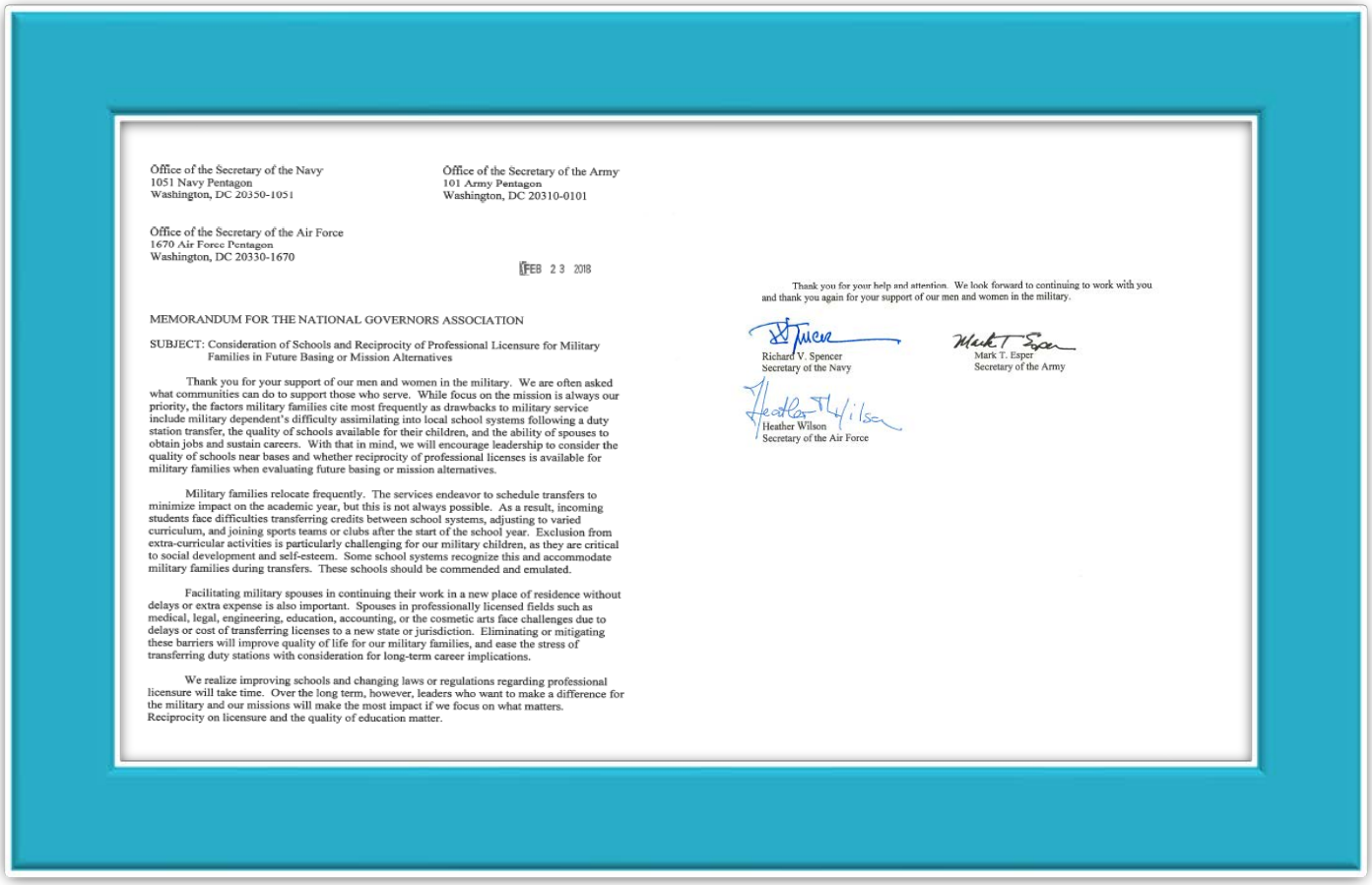
Support of Military Families

The Roosevelt Group and Matrix Design Group were selected by Enterprise Florida to identify opportunities to improve the State's ratings in Air Force's Support of Military Families' assessment for schools and reciprocity. From the onset, TRG/Matrix strove to fully understand the key issues and concerns of Enterprise Florida and the Florida Department of Education. Throughout the execution of this contract TRG worked with stakeholders throughout the State of Florida to identify concerns and to identify specific recommended changes and opportunities to increase the State's competitiveness moving forward. Additionally, the TRG/Matrix team thoroughly evaluated the Air Force's scoring model and compared Florida results with other states/installations across the United States as well as other nationally recognized scoring models.

TRG has also created a detailed engagement plan for Enterprise Florida and State leaders to meet with the Department of the Air Force and the Office of Secretary of Defense to highlight key concerns and to propose improvements and recommendations to the current scoring methodology. This report is being submitted in advance of the actual engagements due to other pressing priorities in Florida, however, TRG will continue to work to support these engagements including identifying specific individuals and their respective roles, schedule meetings, prepare talking points, prepare all Florida participants, and prepare briefings/leave behinds for each of the meetings.

Background:

On February 23, 2018 a letter was signed, by the Secretaries of each of the military departments, and sent to the National Governor's Association. The letter highlighted the need for quality education and opportunities for spousal employment for military families that move every few years and stated that they would be including these factors in future basing decisions. This letter was a call to action across the United States and provided defense communities a clear understanding on how they can help. Ultimately, the quality of education and ability for spousal employment is a retention issue and impacts family resilience.



Office of the Secretary of the Navy
1051 Navy Pentagon
Washington, DC 20350-1051

Office of the Secretary of the Army
101 Army Pentagon
Washington, DC 20310-0101

Office of the Secretary of the Air Force
1670 Air Force Pentagon
Washington, DC 20330-1670

FEB 23 2018

MEMORANDUM FOR THE NATIONAL GOVERNORS ASSOCIATION

SUBJECT: Consideration of Schools and Reciprocity of Professional Licensure for Military Families in Future Basing or Mission Alternatives

Thank you for your support of our men and women in the military. We are often asked what communities can do to support those who serve. While focus on the mission is always our priority, the factors military families cite most frequently as drawbacks to military service include military dependent's difficulty assimilating into local school systems following a duty station transfer, the quality of schools available for their children, and the ability of spouses to obtain jobs and sustain careers. With that in mind, we will encourage leadership to consider the quality of schools near bases and whether reciprocity of professional licenses is available for military families when evaluating future basing or mission alternatives.

Military families relocate frequently. The services endeavor to schedule transfers to minimize impact on the academic year, but this is not always possible. As a result, incoming students face difficulties transferring credits between school systems, adjusting to varied curriculum, and joining sports teams or clubs after the start of the school year. Exclusion from extra-curricular activities is particularly challenging for our military children, as they are critical to social development and self-esteem. Some school systems recognize this and accommodate military families during transfers. These schools should be commended and emulated.

Facilitating military spouses in continuing their work in a new place of residence without delays or extra expense is also important. Spouses in professionally licensed fields such as medical, legal, engineering, education, accounting, or the cosmetic arts face challenges due to delays or cost of transferring licenses to a new state or jurisdiction. Eliminating or mitigating these barriers will improve quality of life for our military families, and ease the stress of transferring duty stations with consideration for long-term career implications.

We realize improving schools and changing laws or regulations regarding professional licensure will take time. Over the long term, however, leaders who want to make a difference for the military and our missions will make the most impact if we focus on what matters. Reciprocity on licensure and the quality of education matter.

Thank you for your help and attention. We look forward to continuing to work with you and thank you again for your support of our men and women in the military.

Richard V. Spencer
Secretary of the Navy

Mark T. Esper
Secretary of the Army

Heather Wilson
Secretary of the Air Force

The Air Force Methodology:

The Department of the Air Force took this direction one step further than the other military services and launched the Support of Military Families program. The Department of the Air Force worked

with many stakeholders, both inside and outside the Department and developed a quantitative framework to measure the quality of education across 157 installations in the United States.

Feb 2018 letter by Service Secretaries has driven significant interest and action across the state

- Air Force Support of Military Families is an extraordinary effort to compare school opportunities for military families across the nation
- Appreciate the Air Force's openness as this analytical framework was being developed and for your willingness to continue to speak with our communities throughout the State
- However, we remain deeply concerned the model provides a misleading picture of the state of Education in Florida
- The scorecard, as currently provided to the public, does not provide enough information for us to decide how to move the needle. Where do we invest our next $??
- Use of MHAs in the analysis do not necessarily represent the districts where the students are attending school--especially for National Guard bases.
- Giving the communities measurable objectives to achieve would help us to advocate for and target resources…similar to reciprocity

The framework the Air Force utilized was based on three factors – academic performance, school climate and service offerings. Simply, the Air Force then analyzed school districts, against this analytical rubric, within the Military Housing Areas for each military installations, pulled data from publicly available sources and then scored each installation against the model (more details on the model are contained in the appendix). Each of the installations was then racked-and-stacked from lowest to highest and also broken down into top, middle and bottom thirds. Across the board, all military installations in Florida did poorly. However, the State of Florida does not believe this is an accurate representation of the quality of schools across the state.

Florida Education Reforms:

The Florida legislature has passed numerous major education reforms over the last four years that support military families. In addition, the State of Florida ranks high in many categories in national testing.

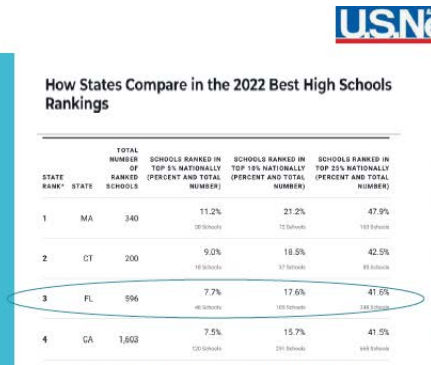
Florida takes the issue of education of military children very seriously and performance reflects that

- Florida legislature has passed major education reforms over last 4 years
  - FL Senate Bill 662 provides military families the ability to register for school before they have an established address, Active Duty Service members only need to present hardcopy PCS Orders.
  - FL House Bill 7045, passed in 2021. The Education Options branch provides the option for K-12 students to attend a participating private school at no cost. Public education funding is portable in the state of Florida for military families (and police officers) and can be used to offset the cost of private schools. and the Step Up For Students Scholarship. Family empowerment scholarship provides designed to offer families of students with disabilities, as young as 3 years of age, access to additional education options.
- Florida ranks first in the nation for participation in Advanced Placement courses during high school and fourth in the nation for performance on AP exams.
- Florida's 4th grade students outperform the national average in both reading and math.
- Florida's 4th grade economically disadvantaged students are also performing higher than the nation in both reading and math. Florida leads the nation with no-cost school choice for all families to take advantage of for students in grades kindergarten through grade 12.
- School choice! Military families get to decide what educational environment they want their children to attend...public, charter, private.

Florida Education Rankings:

The chart below, excerpted from US News and World Report, establishes Florida as one of the top states in the nation for high schools, K-12 and #1 in the nation for higher education.

The State of Florida consistently ranks high in education



https://www.usnews.com/education/best-high-schools/articles/how-states-compare

https://www.usnews.com/news/best-states/rankings/education

US News and World Report recently rated New Jersey as the top state for education. It's followed by Massachusetts, Florida, Washington and Colorado to round out the top five. Florida also was designated as third best high schools in the nation.

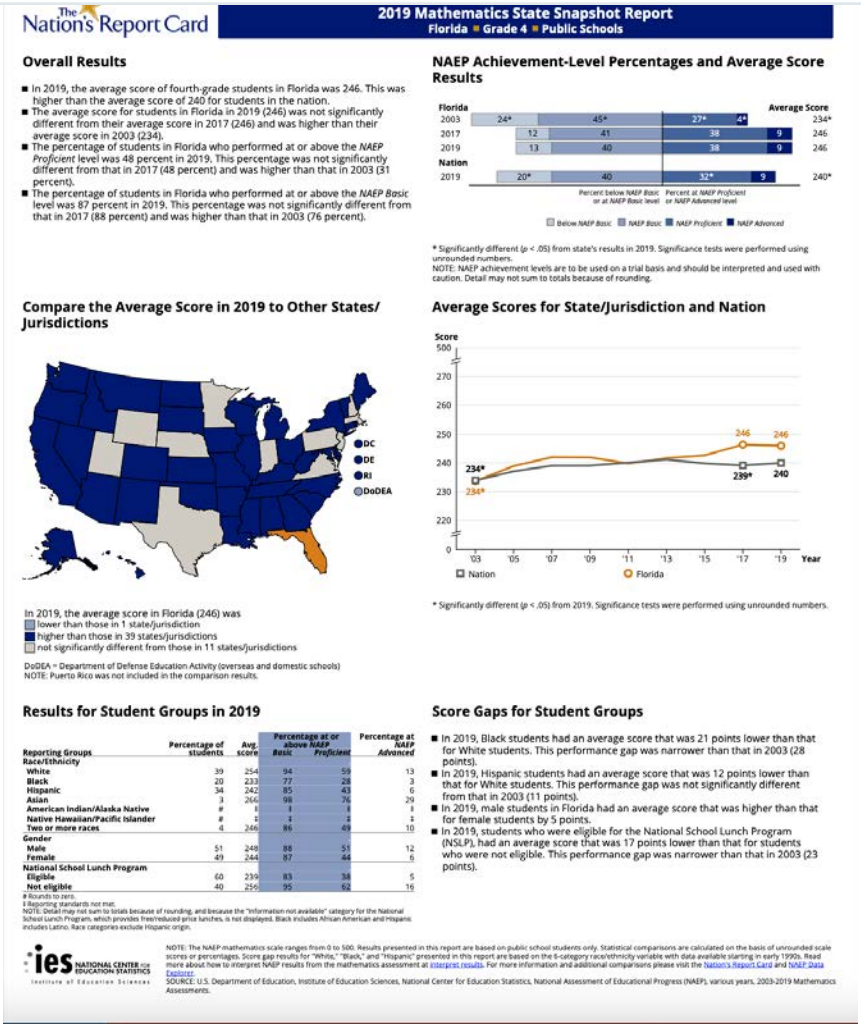
US News and World Report recently rated New Jersey as the top state for education. It's followed by Massachusetts, **Florida**, Washington and Colorado to rough out the top five. **Florida** also was designated as third best high schools in the nation.

The State of Florida consistently ranks high in education

NAEP, also known as The Nation's Report Card, has been providing analysis and results to improve education policy and practice since 1969 and is a congressionally mandated program overseen and administered by the National Center for Education Statistics (NCES), within the U.S. Department of Education and the Institute of Education Sciences. The National Assessment of Education Progress (NAEP) also demonstrates the State of Florida ranking above the national average. Education Week's most recent analysis rated Florida as 3 in the nation.

IES NCES National Center for Education Statistics

NAEP NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS



The Nation's Report Card

2019 Mathematics State Snapshot Report
Florida ■ Grade 4 ■ Public Schools

Overall Results

- In 2019, the average score of fourth-grade students in Florida was 246. This was higher than the average score of 240 for students in the nation.
- The average score for students in Florida in 2019 (246) was not significantly different from their average score in 2017 (246) and was higher than their average score in 2003 (234).
- The percentage of students in Florida who performed at or above the *NAEP Proficient* level was 48 percent in 2019. This percentage was not significantly different from that in 2017 (48 percent) and was higher than that in 2003 (31 percent).
- The percentage of students in Florida who performed at or above the *NAEP Basic* level was 87 percent in 2019. This percentage was not significantly different from that in 2017 (88 percent) and was higher than that in 2003 (76 percent).

NAEP Achievement-Level Percentages and Average Score Results

* Significantly different ($p < .05$) from state's results in 2019. Significance tests were performed using unrounded numbers.
NOTE: NAEP achievement levels are to be used on a trial basis and should be interpreted and used with caution. Detail may not sum to totals because of rounding.

Compare the Average Score in 2019 to Other States/ Jurisdictions

In 2019, the average score in Florida (246) was
lower than those in 1 state/jurisdiction
higher than those in 39 states/jurisdictions
not significantly different from those in 11 states/jurisdictions

DoDEA = Department of Defense Education Activity (overseas and domestic schools)
NOTE: Puerto Rico was not included in the comparison results.

Average Scores for State/Jurisdiction and Nation

* Significantly different ($p < .05$) from 2019. Significance tests were performed using unrounded numbers.

Results for Student Groups in 2019

Reporting Groups	Percentage of students	Avg. score	Percentage at or above NAEP Basic	Percentage at or above NAEP Proficient	Percentage at NAEP Advanced
Race/ethnicity					
White	39	254	94	59	13
Black	20	233	77	28	3
Hispanic	34	242	85	43	6
Asian	3	264	98	76	29
American Indian/Alaska Native	#	‡	‡	‡	‡
Native Hawaiian/Pacific Islander	#	‡	‡	‡	‡
Two or more races	4	246	86	49	10
Gender					
Male	51	248	88	51	12
Female	49	244	87	44	6
National School Lunch Program					
Eligible	60	239	83	38	5
Not eligible	40	256	95	62	16

Rounds to zero.
‡ Reporting standards not met.
NOTE: Detail may not sum to totals because of rounding, and because the "Information not available" category for the National School Lunch Program, which provides free/reduced-price lunches, is not displayed. Black includes African American and Hispanic includes Latino. Race categories exclude Hispanic origin.

Score Gaps for Student Groups

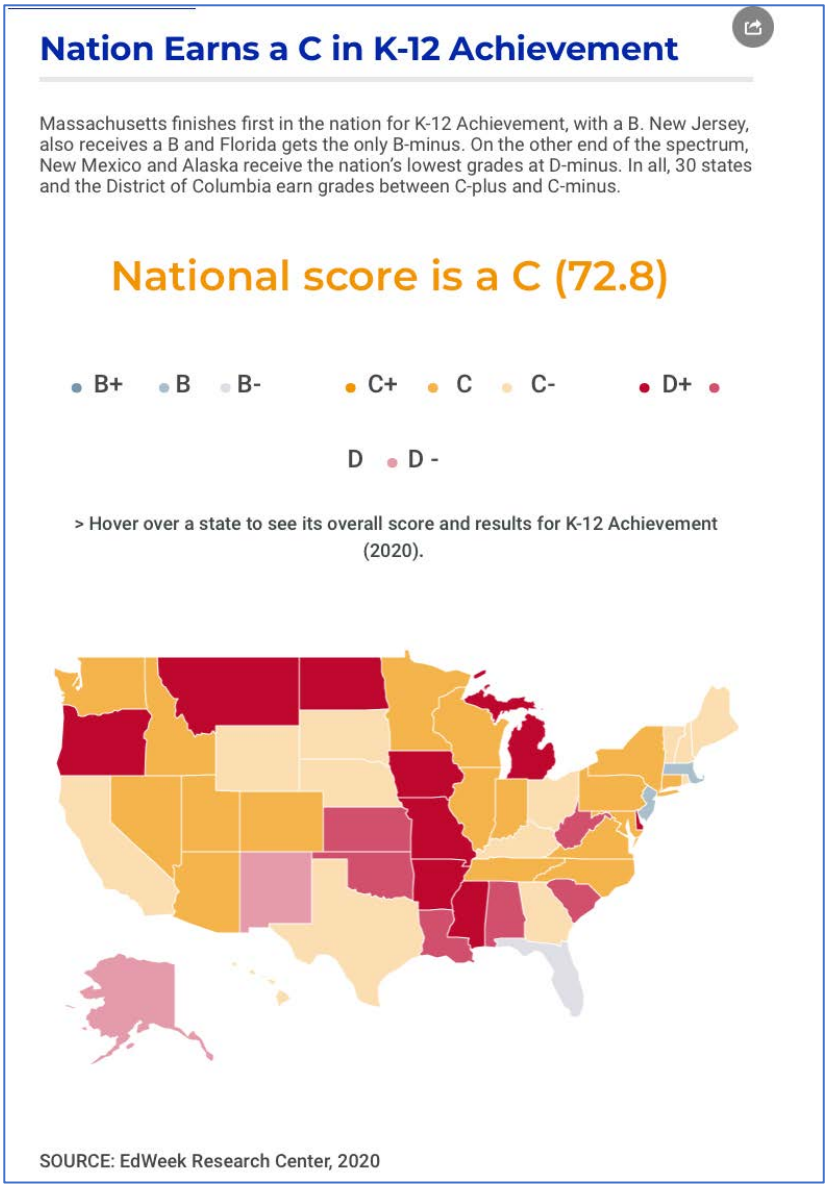
- In 2019, Black students had an average score that was 21 points lower than that for White students. This performance gap was narrower than that in 2003 (28 points).
- In 2019, Hispanic students had an average score that was 12 points lower than that for White students. This performance gap was not significantly different from that in 2003 (11 points).
- In 2019, male students in Florida had an average score that was higher than that for female students by 5 points.
- In 2019, students who were eligible for the National School Lunch Program (NSLP), had an average score that was 17 points lower than that for students who were not eligible. This performance gap was narrower than that in 2003 (23 points).

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

NOTE: The NAEP mathematics scale ranges from 0 to 500. Results presented in this report are based on public school students only. Statistical comparisons are calculated on the basis of unrounded scale scores or percentages. Score gap results for "White," "Black," and "Hispanic" presented in this report are based on the 6-category race/ethnicity variable with data available starting in early 1990s. Read more about how to interpret NAEP results from the mathematics assessment at interpret results. For more information and additional comparisons please visit the Nation's Report Card and NAEP Data Explorer.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 2003-2019 Mathematics Assessments.

https://nces.ed.gov/nationsreportcard/

THE Roosevelt GROUP Matrix ABLE OPERATIONS ECONOMIC & WORKFORCE ANALYTICS



https://www.edweek.org/policy-politics/nation-gets-a-c-minus-on-k-12-achievement/2020/09

The most recent Air Force Support of Military Families results consistently ranked all Florida military installations near the bottom of the scale. The Roosevelt Group and Matrix were asked to delve into these results and try to determine why there is such a significant difference from other models and to provide feedback to the Air Force.

But…the Air Force analysis has Florida trending near the bottom for all 9 Air Force installations… why the disconnect?

Analytical Framework:

The Air Force recognized up-front the challenges associated with trying to develop a model as complex as this and has stated "The Support of Military Families program team is listening, and we consistently rely on feedback to understand location specific concerns as part of our evaluation of the current frameworks and as we consider any specific adjustments for the future." Enterprise Florida and the State Department of Education have extensively reviewed the results of the SOMF scoring and plan to address with senior leaders in both the Department of the Air Force as well as the Office of the Secretary of Defense for Personnel and Readiness. The remainder of this report and its associated appendix will capture some of the key concerns and requests for the Department of the Air Force.

The following chart highlights the Air Force's assessment methodology against the US News and World Report factors. Our review determined the US News and World Report Factors provide a much more in-depth review of K-12 and breaks it down into K-8 and high school. In fact, they recognize that simply looking at graduation rates for high school was not an adequate analysis of the quality of the education at that level. The Air Force SOMF Assessment seemingly pays little attention to high school – other than graduation rates. Whereas, US News and World includes college readiness, college curriculum breath, and proficiency and performance in both math and reading. The SOMF does not measure any of those factors. Further, the weighting for graduation rate is a 30% for AF while US News and World is 10%.

A final observation on the Air Force methodology is the use of both input and output-based data for their analysis. The US News and World Report analysis was all output based and on actual performance.

Learning rate was derived from Stanford Education Data Archive (SEDA) 30% of overall AF score

Air Force SOMF Assessment Weights

Academic Performance	Grad Rates	30%
	Learning Rate	30%
School Climate	Chronic Absenteeism	10%
	Suspension Rate	10%
Service Offering	Free and/or Universal Pre-K	4%
	Student to Counselor Ratio	4%
	Student to Mental Health Support Ratio	4%
	Student to Nurse Ratio	4%
	Student to Teacher Ratio	4%

Input and output based factors and only measure to Grade 8

US News and World Report Factors

High School

College Readiness	30%
College Curriculum Breadth	10%
Math and Reading Proficiency	20%
Math and Reading Performance	20%
Underserved Student Performance	10%
Graduation Rates	10%

K-8

Math and Reading Proficiency	50%
Math and Reading Performance	50%

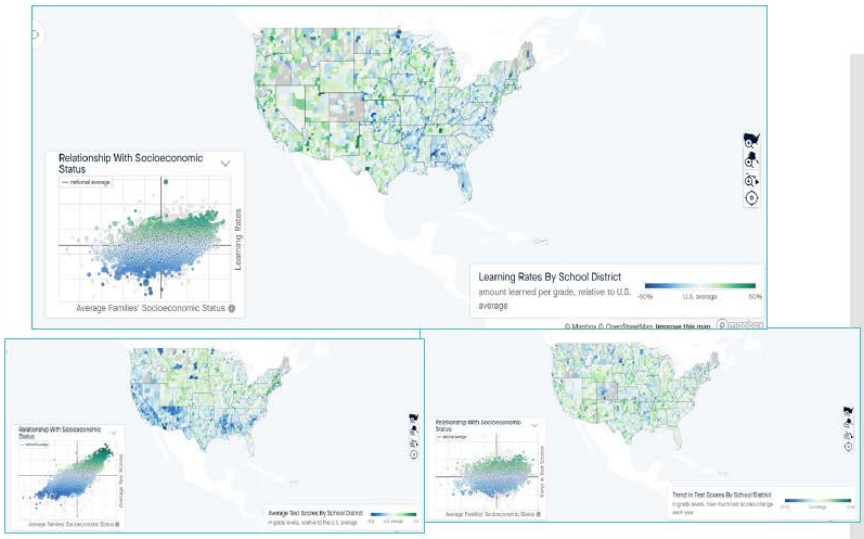
All output based factors and measure Pre-K through 12

A significant concern from the analysis and the State of Florida is the Air Force's use of the SEDA data. A full 30% of the score is based on a learning rate -- the rate of growth between 3rd and 8th grade. Florida consistently scores about the national average in testing in both math and reading but does worse than some of the nation in the rate of growth. However, our evaluation determined other states that ended up in the "green" area of the Air Force analysis performed below average in both math and reading but their rate of growth was higher. From a parent perspective which school would I want my child to attend? The one with better scores or the one that still had scores below average but improved at a higher rate?

The top map on the following chart depicts the data the AF used as 30% of their overall score. Note that most of Florida was colored blue - which is below the national average. However, looking at the bottom two charts – Florida is predominantly green. Test scores and trend in test scores in Florida both exceeded the national average. Consequently, it is illogical that Florida rated as poorly as it did in the Air Force model.

Stanford Education Data Archive (SEDA) Results

Florida is below average in learning rates but above average in both average test scores and trend in test scores



Obtained from SEDA
https://edopportunity.org/explorer/#/map/none/districts/coh/ses/all/2.64/30.31/-94.15/

Concerns and Recommendations:

Additional thoughts on our concerns over the use of the SEDA model are contained in the following chart as well as the following discussion on the methodological review.

General Methodological Concerns

SEDA was designed for research purposes and not business decisions, for which the AF is currently using it.

- SEDA data, which constitutes 30% of scoring, utilizes state-level NAEP performance to normalize district/school level standardized testing performance. Does AF want to measure learning growth or what students actually know in Grade 8?? **Using learning growth measures rate of improvement vs actual performance.**
- The academic community has yet to agree on the validity and/or limitations of SEDA results.
- Too many differing state-specific challenges, practices, and objectives to develop a meaningful assessment across states; in addition to differences in standardized testing, state policies differ in staffing, funding, and certification requirements, leading to comparison of uncommon measures.
- Assessment does not control for socio-economic factors, which could disproportionally benefit communities with higher socio-economic demographics

Defense communities understand the factors but are having difficulty in determining how/where to move the needle to make improvements.

There continues to be community concerns on the Air Force's intention with using this analysis as part of their strategic basing process. If used during the site survey phase, and communities are

given the opportunity to provide input, the concerns are somewhat mitigated. However, if the AF uses it at the start of the process, and it is a factor in deciding the candidate bases, a base may well be eliminated from further look if they did not score well in the SOMF analysis.

Additionally, please note the school districts that were included as part of the SOMF analysis for Jacksonville ANG and Homestead Air Reserve Base. This chart clearly demonstrates that the location the reserve member decides to live is not based on schools, but other factors such as full-time employment. This is different methodology to active duty which uses the MHA as the capture area.

Concerns on how it will be used in Strategic Basing Process

- If used at the beginning of the Strategic Basing Process a base may never make it to the "candidate" selection phase.
  - 25 out of 125 points will keep lower bases from even being considered
  - AF has stated that states/communities will have time during site surveys to "set the record straight" but in fact wont make it to the list
- Air National Guard members are not subject to moves like their active duty counterparts. They often travel long distances to their Guard positions so should schools even be a consideration in ANG basing actions?

The following chart is provided for context. Communities are receiving multiple messages from DoD on schools. DoD's endorsement of the Purple Star School designation and input from the Defense State Liaison officials as well as the separate AF SOMF analysis is confusing to defense communities.

OSD tracks Purple Star Schools Program

Florida passed!

Different measures cause confusion

Key Message
Military children experience many challenges as they relocate to new schools due to a parent's change in duty station. By establishing statewide Purple Star Schools programs, states can encourage local education agencies to implement practices that assist military children with transitions/deployments and also recognize military service and civic responsibility.

Discussion Points
Military children face issues such as gaps and overlaps in curriculum, different graduation requirements and course-placement disruption, as well as difficulties related to socially and emotionally connecting with new schools and communities.
State-sponsored Purple Star Schools programs primarily distinguish schools as military friendly once they demonstrate a major commitment to students and families connected to our nation's military.

NOTE: Florida DoE Purple Star Schools of Distinction Criteria. Florida has a much more rigorous program. For instance, staff development is for the entire staff not just the Military Point of Contact (MPOC), 5% of Controlled Open Enrollment Seats must be reserved for Military Affiliated Student, schools must have a student-led (or peer-to-peer) transition support program, and schools must participate in three yearly activities supporting the military community as opposed to one.

Through the Purple Star program, states have also recognized schools with few military families that have used the criteria to emphasize military service and recognize veterans and active members of the National Guard and reserves in their community. Schools have established programs that:

- Recognize the accomplishments of veterans and active members
- Sponsor special events recognizing military service
- Celebrate class members who commit to serving in the military

Source: MilitaryOneSource

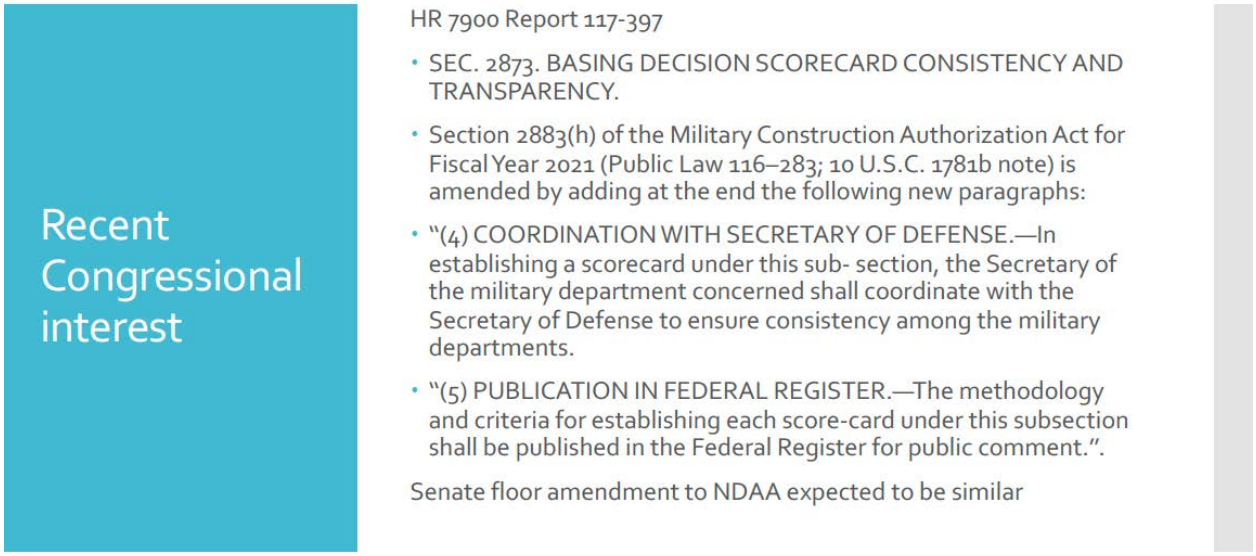
https://statepolicy.militaryonesource.mil/status-tracker/purple-star-schools-program

Most importantly, we suggest DoD to look to measure what parents are looking for when choosing a school for their children. What are those critical offerings that a school/district must have to support the military family and their students?

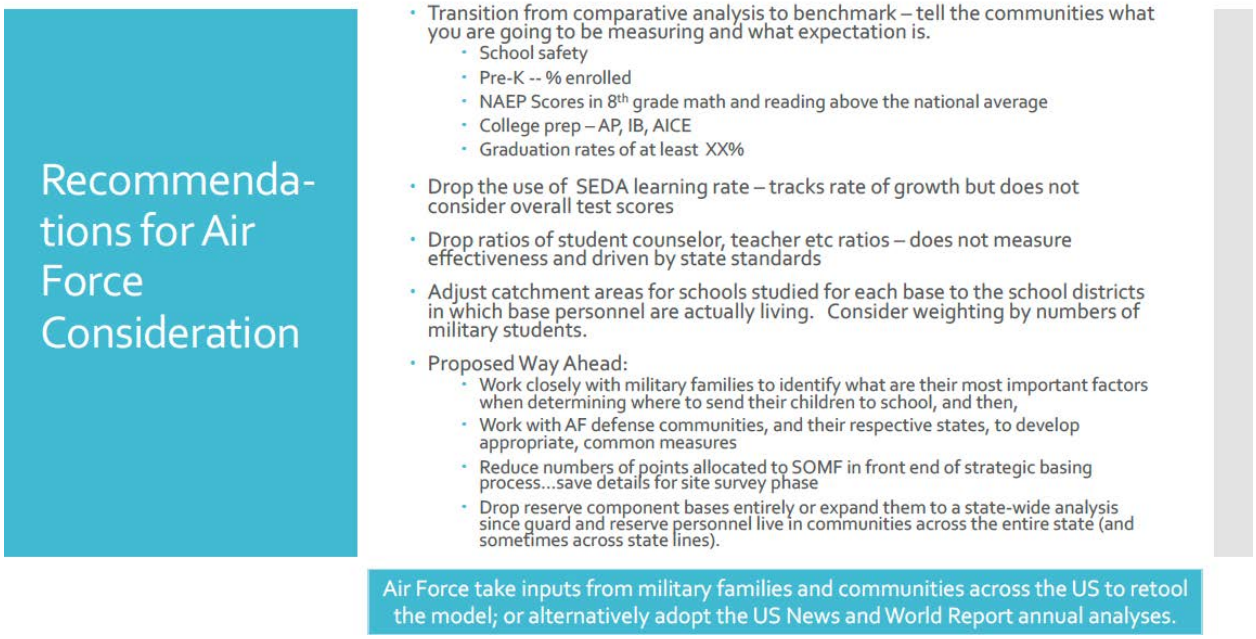
What are families looking for when moving to a new duty station?

- School safety
- State/local level ranking of schools
- Individual offerings for career and technical tracks
- Sports offerings
- ROTC program
- Extracurricular activities – orchestra, military support programs
- Access to before/after school care
- Open enrollment opportunities
- Quality of virtual school
- AP/IB/ACE program and ability to dual enroll in college
- Home education curriculum

There has been significant push back and concerns across the nation as to the Air Force's scoring methodology. During the FY23 NDAA action by the House of Representatives the following language was included that would require DoD to coordinate across the military departments to ensure consistency and to publish the score card methodology in the Federal Register.

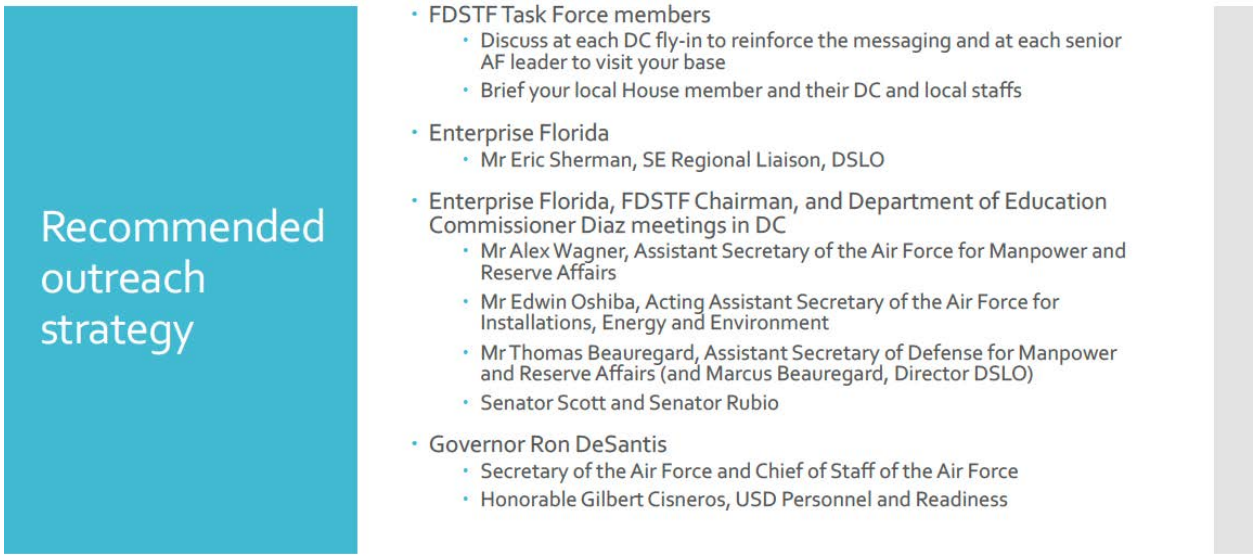


The following chart summarizes the State of Florida's recommendations to the Air Force for consideration. In one sentence: **transition to a bench mark – tell the communities what the Air Force expects of the schools/districts and measure that.** The Air Force does not tell 1/3 of their military members they are in the bottom third but that is exactly what the AF is doing to their defense communities. Why wouldn't the Air Force want to raise all their locations to a standard versus forcing each location into a top, middle or bottom third bucket. This is very disheartening to communities that work very hard each day to support their military installations. As mentioned at the beginning of this report, the February 23, 2018 letter signed by all the Service Secretaries has created an awareness, and commitment, across the United States by defense communities to improve their schools.



Outreach Strategy:

The final chart briefly highlights the organizations and people the State of Florida should meet with about their concerns on the SOMF analysis.



The following pages go into greater details on the Methodological Review of the SOMF analysis.

Support of Military Families Methodological Review

Introduction

The Florida Defense Support Task Force commissioned The Roosevelt Group (and Matrix Design Group through sub consulting) to review and critique the Air Force's (AF) Support of Military Families community assessment, specifically the public education grading rubric and methodology – ultimately leading to an AF engagement strategy to influence the Department on utilizing a more effective assessment. Matrix Design Group was tasked with the following:

1. As the AF did not release community scores, compile data used in the assessment to provide the actual raw score rather than the red, yellow, green marks assigned by the AF.
2. Review the assessment's methodology and provide an assessment of each component.
3. Develop recommendations for either the State, local school districts, or the AF to help improve scores or the assessments methodology.

Support of Military Families Public Education Methodology

PURPOSE

Local support for military members and families who reside on and around our installations is an important factor in total force readiness. The strategic importance of this initiative is to ensure locations where we place our military members and their families provide the capabilities necessary to enhance our military family readiness and improves member retention. The Department of the Air Force is dedicated to bringing awareness to, and mitigating, factors that negatively affect readiness and retention for military members and their families as they transition from one duty location to the next.[1]

Background: The decision to continue military service is influenced by public education opportunities for military children. To address this issue, the Secretaries of the Army, Navy and Air Force informed the National Governor's Association that...

"Eliminating or mitigating these barriers will improve quality of life for our military families, and ease the stress of transferring duty stations with consideration for long-term career implications. We realize improving schools... will take time. Over the long term, however, leaders who want to make a difference for the military and our missions will make the most impact if we focus on what matters."

In partnership with policy and industry experts, and key stakeholders, the Department of the Air Force developed an analytic framework using quantitative criteria to assess public education. This methodology assesses the school districts' support for the unique needs of military children within military housing areas surrounding an installation.[2]

[1] https://www.af.mil/Portals/1/documents/2021SAF/09_Sept/External_CASH_single_map_file_v4.2.pdf
[2] Ibid

Framework: Careful consideration used to reduce the impact of socioeconomic factors while selecting criteria, and all data was obtained from publicly available and reputable sources.

Academic Performance: The most important area, this measures student learning and successful program completion.

School Climate: Captures whether the schools provide an environment supportive of academic learning.

Service Offerings: Includes programs and staff designed to ease transitions and provide emotional and academic support to students.

SCORING RUBRIC

Air Force Assessment Weights

Academic Performance	Grad Rates	30%
	Learning Rate	30%
School Climate	Chronic Absenteeism	10%
	Suspension Rate	10%
Service Offerings	Free and/or Universal Pre-K	4%
	Student to Counselor Ratio	4%
	Student to Mental Health Support Ratio	4%
	Student to Nurse Ratio	4%
	Student to Teacher Ratio	4%

DATA SOURCES

Stanford Education Data Archive harnesses data from the U.S. Department of Education EDFacts data system and a number of other publicly available data files to aid scholars, policymakers, and educators. The information includes measures of academic opportunity and gaps based on socioeconomic status.

U.S. Department of Education: EDFacts Graduation Rates (District and School Level) EDFacts is a U.S. Department of Education initiative to collect, analyze, and promote the use of high-quality, pre-kindergarten through grade12 data.

U.S. Department of Education: Civil Rights Data Collection (CRDC) CRDC gathers information on student enrollment, education programs, and school services, broken down by race, sex, English proficiency, and disability. The data is collected biennially from every public school in the United States.

Universal Methodological Concerns

Several universal methodological issues were uncovered during data analysis, literature review, and stakeholder discussions. Many methodological concerns that have already been submitted to the AF are excluded here.

1. Use of differing years for source data does not accurately depict much about a given year's performance. Graduation rates were from 2018 - 2019, while school climate, service offerings, and enrollment data were from 2017 -2018.
2. As data lags several years, school districts and states have little ability to influence their performance until several iterations later, potentially losing out on missions and investment. The assessment is backward looking, when strategic basing decisions are forward looking by definition.
3. While the AF considered the differences in state standardized testing criteria as an obstacle to comparing standardized test scores across state lines (as with the Learning Rate metric), they did not control for these differences for the other metrics. As with standardized testing, states regulate and fund mental health support, nursing and teaching loads, and student counselor requirements differently, making their comparison across state lines difficult, if not impossible.
4. Although the AF claims it attempted to reduce the impact of socioeconomic factors while selecting criteria, no such control appears to have been done for school climate ratings (suspension and chronic absenteeism). By not controlling for these factors, AF communities with higher socioeconomic demographics may benefit disproportionately.
5. A handful of mistakes were caught in the 2021 Support of Military Family update, such as miscolored communities, wrong school district names, and missing school districts. While not direct evidence of analytical / calculation error, it does raise concerns.
6. For the learning rate metric, the use of National Assessment of Educational Progress (NAEP) performance data, which is a national standardized test given to a sample of students in every state, to normalize state-level standardized testing scores at the district level is of concern. These normalized scores are produced by the Sandford Center for Education Policy Analysis and are referred to as the Stanford Education Data Archive (SEDA). The academic community has raised several concerns with attempting to normalize testing scores that, due to federalism and state control over education policy, are inherently uncommon measures. A detailed discussion of the academia's view of the SEDA data are provided in the literature review at the end of this report, but while great progress has been made in the statistical procedures used in the analysis, the academic community is not in agreement over the validity of SEDA's outputs.
7. As with all statistical procedures, the outputs are estimates with varying margins of error. It is unclear if the AF identified acceptable margins of error for use in the assessment, but generally

researchers identify acceptable ranges of error and whether or not an output achieved the minimum range. Not doing so raises concern over the validity of results.

8. As the NAEP scores used to normalize school level testing scores is generated from state-level assessment, it is unclear that normalized scores truly reflect a district's performance or merely infers district-level performance from state-level NAEP performance.
9. Finally, SEDA was designed for research purposes and not for making multibillion-dollar business decisions that could adversely impact communities that receive poor marks.

Academic Performance

Educational metrics that establish a foundation for college and or career readiness. Assesses student learning and successful high school graduation.

Graduation Rate

1. Four-year graduation rate of all eligible students within the school district
2. Typically, published annually in January via the Department of Education EDFacts Data System

Air Force Community	Graduation Rates
Jacksonville Air National Guard Base, Florida	89.55
Duke Field, Florida	88.23
Eglin Air Force Base, Florida	88.23
Hurlburt Field, Florida	88.23
Cape Canaveral Space Force Station, Florida	88.00
Patrick Space Force Base, Florida	88.00
MacDill Air Force Base, Florida	86.96 / 86.98
Homestead Air Reserve Base, Florida	86.58
Tyndall Air Force Base, Florida	83.04

Source: https://www2.ed.gov/about/inits/ed/edfacts/data-files/index.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Presented above, all but two Florida AF communities received a yellow marking meaning they fell within the middle 1/3 of AF communities. These middle-tiered communities graduated between an 86.58% (Homestead ARB) and 88.23% (Eglin, Hurlburt, and Duke). The two remaining Florida communities - Jacksonville ANGB and Tyndall AFB - received green and red marks, respectively. Jacksonville ANGB graduated 89.55% and Tyndall graduated 83.04% of their students. The

standard deviation for all Florida AF communities was 1.8, while the standard deviation for just those middle-tiered communities was .7. As Tyndall AFB had the lowest graduation rate at 83.04%, or nearly 3.54 point lower than the lowest middle-tiered community (Homestead ARB), Tyndall was a clear outlier, and when Tyndall is removed from the analysis, the standard deviation is .9.

Assuming Jacksonville ANGB is at the lower end of the top-tiered AF communities, five Florida AF communities were within 1.6 percentage points of receiving green marks, while two more were within 3 percentage points. While MacDill, Homestead, and Tyndall, will likely have challenges overcoming their middle and lower tiered status, Duke, Eglin, Hurlburt, Cape Canaveral and Patrick AF communities are likely at the upper-end of the middle-tier grouping.

ISSUES WITH COMPARING GRADUATION RATES ACROSS STATES

The use of graduation rates as a metric within the assessment poses similar challenges as most every metric used – graduation requirements differ across states and therefore are not an apples-to-apples comparison. For example, 47 states have a defined state graduation minimum requirement, while Massachusetts, Pennsylvania, and Colorado have permitted autonomy to the State's individual school district to determine graduation requirements. Even those state's that have state defined requirements, standards can range considerably from state-to-state. For example, California has a minimum requirement of 13 total credits, while Florida has a minimum of 24 total credits.[3] To make matters more complicated, local California school districts may add its own requirements.[4] While additional complications exist, differing minimum graduation requirements is sufficient to determine that utilizing raw graduation rates in a comparative assessment cannot produce a meaningful analysis, particularly if there is no attempt to control for state differences. More to the point, as the AF uses graduation rates as a proxy for college readiness, research suggest other factors are much better predictors of college success, such as high school GPAs rather than standardized tests and graduation rates. [5]

School Climate

Indicators of a safe educational environment and its contribution to academic learning.

Chronic Absenteeism

- Rate of students that have chronic absenteeism, as defined by missing at least 15 days of

[3] https://nces.ed.gov/programs/statereform/tab3_3-2020.asp
[4] https://www.cde.ca.gov/ci/gs/hs/hsgrfaq.asp
[5] https://www.forbes.com/sites/nickmorrison/2020/01/29/its-gpas-not-standardized-tests-that-predict-college-success/?sh=4aaa8b1d32bd

schools in a given school year

- Reported once every other year to Department of Education - Civil Rights Data Collection

Installation	Chronic Absenteeism
Robins AFB, Georgia	13.68
Cape Canaveral Space Force Station, Florida	15.43
Patrick Space Force Base, Florida	15.43
Duke Field, Florida	20.38
Eglin Air Force Base, Florida	20.38
Hurlburt Field, Florida	20.38
Homestead Air Reserve Base, Florida	20.42
MacDill Air Force Base, Florida	21.85 / 21.82
Jacksonville Air National Guard Base, Florida	26.65
Tyndall Air Force Base, Florida	32.23
Standard Deviation	**5.53**

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Six of the nine Florida AF communities received yellow marks with a standard deviation of 2.71.[6] The remaining three communities received red marks with a standard deviation 5.19. Robins AFB, in Georgia, received a green mark and is used as the benchmark, upper-tier comparison community. When all 10 communities are included, the standard deviation more than doubles to 5.53. In totality, the dispersion of observations, assuming Robins is the benchmark upper-tier score, is significant with the poles driving much of the standard deviation. Cape Canaveral and Patrick SFBs (Brevard County School District) is the only top performing district of the Florida communities and is less than two points from a green mark.

ISSUES WITH COMPARING CHRONIC ABSENTEEISM RATES ACROSS STATES

There appears to be no data quality issues with using chronic absenteeism as a metric within the AF's assessment of public schools. The federal government defines chronic absenteeism as missing at least 15 days of schools in a year. While certain local factors could impact this rate,

[6] The standard deviation within these observations is relatively small due, in part, to five of them being located in the same school districts – Cape Canaveral and Patrick SFB in Brevard and Duke, Eglin, and Hurlburt in Okaloosa. However, observations are also tightly clustered as the standard deviation only increases to 2.87 (from 2.71) when duplicates are removed.

such as hurricanes or pandemics that could materially impact school attendance, given the definition is standard across school districts (and states), this metric seems to cause little data quality concerns. However, some stakeholders raised concerns that chronic absenteeism is determined more by family / household life rather than the quality of a school district. As such, comparing school districts without controlling for socioeconomic factors may distort assessment outcomes and thus may favor school districts with higher socioeconomic demographics, which is a methodological concern.

Suspension Rate

• Rate of students from grades Pre-Kindergarten through 12th grade with and without disabilities who received at least one suspension (in and/or out of school)

• Reported once every other year to Department of Education - Civil Rights Data Collection

Installation	Suspension Rate
Eleison Air Force Base, Alaska	7.43
Homestead Air Reserve Base, Florida	8.68
Duke Field, Florida	12.56
Eglin Air Force Base, Florida	12.56
Hurlburt Field, Florida	12.56
Cape Canaveral Space Force Station, Florida	13.07
Patrick Space Force Base, Florida	13.07
MacDill Air Force Base, Florida	13.19 / 12.79
Jacksonville Air National Guard Base, Florida	13.63
Tyndall Air Force Base, Florida	20.14
Standard Deviation	**3.35**

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Six of the nine Florida AF communities received yellow marks while three received red. The middle-tiered communities receiving yellow marks had an adjusted standard deviation of 2.4. Using Eleison AFB in Alaska as the upper tier, or green, benchmark community, the 10 observations had a standard deviation of 3.35. Homestead ARB was the highest performing Florida AF community with a suspension rate of 8.68, which is 1.25 points from Eielson's 7.43 green benchmark score. The remaining Florida AF communities significantly trail the upper-tiered communities.

ISSUES WITH COMPARING SUSPENSION RATES ACROSS STATES

Similar to chronic absenteeism, there appears to be no data quality issues with using suspension rates as a metric within the AF's assessment of public schools. The AF utilizes the rate of students from grades Pre-Kindergarten through 12th grade with and without disabilities who received at least one suspension (in and/or out of school) in their assessment. Given the definition is standard across school districts (and states), this metric seems to cause little data quality concerns. However, as with chronic absenteeism, some stakeholders raised concerns that suspensions are likely driven more by family / household life than the quality of a school district. As such, comparing school districts without controlling for socioeconomic status may distort assessment outcomes and thus may favor school districts with higher socioeconomic demographics, which is a methodological concern. Moreover, certain states, such as Florida, set certain disciplinary rules through statutory language and thus create additional methodological concerns as AF communities being assessed may have different disciplinary regulations which would likely impact suspension rates.

Service Offerings

Access to programs and qualified staff providing specialized services.

Free and/or Universal Pre-Kindergarten

No analysis was done for free and/or universal Pre-kindergarten as every Florida AF community received a green mark.

Student to Counselor Ratio

- Ratio of student enrollment to total counselor FTEs
- Reported once every other year to Department of Education - Civil Rights Data Collection

AF Community	Student/ Counselor Ratio	Percentage of Total Enrollment
Forbes Field KS	345.33	.29%
Cape Canaveral Space Force Station, Florida	402.43	.25%
Patrick Space Force Base, Florida	402.43	.25%
Tyndall Air Force Base, Florida	408.84	.24%
MacDill Air Force Base, Florida	458.63	.22%
Jacksonville Air National Guard Base, Florida	471.01	.21%
Homestead Air Reserve Base, Florida	484.60	.21%
Duke Field, Florida	513.88	.19%
Eglin Air Force Base, Florida	513.88	.19%
Hurlburt Field, Florida	513.88	.19%
Standard Deviation	**58.79**	

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Four of the nine Florida AF communities received yellow marks for their student to counselor ratios, with Cape Canaveral and Patrick SFB having the highest ratio of 402.43 to 1, which equated to .25% of the total student enrollment in Brevard County School District. The benchmark, upper-tier AF communities used here is Forbes Field in Kanas. The community received a green mark with a ratio of 345.33 to 1, or .29% of total student enrollment. As a point of comparison, Hurlburt, Eglin, and Duke (Okaloosa County) received red marks and had a student to counselor ratio of

513.88 to 1, or .19% of their total enrollment. When all ten communities are compared, there is a standard deviation of 58.79.

ISSUES WITH COMPARING STUDENT TO COUNSELOR RATIOS ACROSS STATES

Methodological issues arise when comparing student to counselor ratios in school districts across states. The most important is the inability to effectively assess the quality of counseling given. Does a higher student to counselor ratio actually mean student receive better quality counseling? As states control the educational, experiential, and examination requirements for their school counselors, a simple ratio measuring quantity and quality brings in question effectiveness within the AF's assessment criteria. While all 50 states require a graduate level education to become a student counselor, only 7 require a minimum number of graduate level credit hours to be taken in school counseling prior to employment.[7] Twenty-six states require the completion of a supervised, school-based internship or practicum. Finally, only thirty-four states require the passage of one or more standardized examination prior to employment. Because state differ in the requirements to become a student counselor, utilizing a simple ratio that describes quantity with no emphasis on quality is essentially a useless metric that may reward school districts with higher socioeconomic demographics and small enrollment sizes.

[7] https://www.counseling.org/docs/licensure/schoolcounselingregs2011.pd

Student to Mental Health Support Ratio

- Ratio of student enrollment to the sum of total psychologist FTEs and social worker FTEs
- Reported once every other year to Department of Education - Civil Rights Data Collection

AF Community	Student / Mental Health Support Ratio
Forbes Field KS	323.92
Tyndall Air Force Base, Florida	2,504.17
Homestead Air Reserve Base, Florida	3,487.17
Cape Canaveral Space Force Station, Florida	7,273.66
Patrick Space Force Base, Florida	7,273.66
MacDill Air Force Base, Florida	15,316.71 / 1,186.59
Jacksonville Air National Guard Base, Florida	19,777.07
Duke Field, Florida	No data
Eglin Air Force Base, Florida	No data
Hurlburt Field, Florida	No data
Standard Deviation	**N/A**

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Eight of the nine Florida AF communities received red marks for their student to mental health support ratios - MacDill was the only community that received a yellow mark. With respect to MacDill, when the region is adjusted to include Hillsborough County School District, the community's score drops from 15,316.71 to 1,186.59, which make's sense given the community's yellow score. If Hillsborough is not included, the 15,316.71 to 1 ratio receiving a yellow score when other red communities have higher ratio supports the notion that Hillsborough is included in the assessment, contrary to the AF's 2021 Support of Military Families Update.[8] Forbes Field KS is again used as the benchmark, upper-tier community that received a green mark. On the surface, Forbes Field has a much higher ratio at 323.19 to 1, compared to Florida's highest scoring AF

[8] https://www.af.mil/Portals/1/documents/2021SAF/09_Sept/External_CASH_single_map_file_v4.2.pdf

community at 1,186.59 to 1. However, as discussed below, Florida mental health practitioners may not be accurately reported given changes in mental health funding over the last several years.

ISSUES WITH COMPARING STUDENT TO MENTAL HEALTH SUPPORT RATIOS ACROSS STATES

The raw data used for this metric is housed by the US Department of Education's Office of Civil Rights but uploaded by individual states and school districts. The AF considers the sum of fulltime school psychologists and social workers reported by school districts relative to total enrollment of that district as constituting the district's level of support for student mental health issues. However, stakeholders reported these data do not tell the whole story. And, as these data are from the 2017/2018 school year, do not accurately reflect the significant funding increases by the State of Florida to provide additional mental health services over the last several years. As such, making basing decisions with data that is nearly five years would not reflect the current conditions.

Student to Nurse Ratio

- Ratio of student enrollment to total Nurse FTEs
- Reported once every other year to Department of Education - Civil Rights Data Collection

AF Community	Student / Nurse Ratio
Forbes Field KS	875.11
Jacksonville Air National Guard Base, Florida	1,896.43
Homestead Air Reserve Base, Florida	2,417.44
Duke Field, Florida	3,166.31
Eglin Air Force Base, Florida	3,166.31
Hurlburt Field, Florida	3,166.31
Tyndall Air Force Base, Florida	3,611.78
MacDill Air Force Base, Florida	8,793.18 / 1,531.23
Cape Canaveral Space Force Station, Florida	36,732.00
Patrick Space Force Base, Florida	36,732.00
Standard Deviation	**14,212.51 / 14,467.43**

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Florida AF communities also struggled with student to nurse ratios with seven out of nine communities receiving a red mark. Jacksonville ANGB and MacDill AFB (if Hillsborough is included) were the communities receiving yellow marks. Forbes Field, Kansas, is used again as the benchmark, upper-tier community that received a green mark. Forbes Field received a score of 875.11 to 1, whereas MacDill, the state's highest performing community received 1,531.23 to 1. As with student to mental health support data, several limitations exist with how school districts report data to the Office of Civil Rights and is discussed below.

ISSUES WITH COMPARING STUDENT TO NURSE RATIOS ACROSS STATES

Similar to mental health support, utilizing raw data from the Office of Civil Rights for nurse FTEs does not tell an accurate story of a school district's commitment to providing health services for its student body. Stakeholders cited that FTE count provided to the Office of Civil Rights only includes nurse that are staffed by the school district. Many school districts across the state contract with county health departments or other nonprofits to provide additional nursing support. As this is district by district decision, comparing across states (and within) does provide for an apples-to-apples comparison. An example of this was found in Bay County – home to Tyndall AFB. The Office of Civil Rights reports the school district only funds six nurse FTEs; however, in discussion with the school district it was discovered the county health department funded an additional 15 nurses and 39 health technicians during the same calendar year, for a total 62 nursing support FTEs. When these positions are considered in the analysis, the AF communities' student to nurse ratios increase significantly from 3,374 to 1 to 453 to 1 which places it well ahead of Forbes Field. While its unknown how many school districts located in AF communities across the nation use a similar model, it is clear that the scoring, and likely rankings, would change if these additional personnel are included. This issue alone renders this metric useless as no meaningful information can be gleaned from this approach.

Student to Teacher Ratio

- Ratio of student enrollment to total teacher FTEs (Full-Time Equivalent)
- Reported once every other year to Department of Education - Civil Rights Data Collection

AF Community	Student / Teacher Ratio	Percentage of Total Enrollment
Forbes Field KS	12.82	7.80%
MacDill Air Force Base, Florida	13.96 / 14.66	7.16% / 6.82%
Cape Canaveral Space Force Station, Florida	16.07	6.22%
Patrick Space Force Base, Florida	16.07	6.22%
Tyndall Air Force Base, Florida	16.16	6.19%
Duke Field, Florida	16.46	6.07%
Eglin Air Force Base, Florida	16.46	6.07%
Hurlburt Field, Florida	16.46	6.07%
Jacksonville Air National Guard Base, Florida	16.92	5.91%
Homestead Air Reserve Base, Florida	17.30	5.78%
Standard Deviation	**1.39 / 1.29**	

Source: https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html
Note: For MacDill AFB, the AF documentation used to calculate scores did not include Hillsborough County School District which is county in which MacDill AFB resides. Based on discussions with stakeholders, it was determined that this is likely an error so calculations were provided.

FLORIDA AIR FORCE COMMUNITY SCORES

Five Florida AF communities received yellow marks with the remaining four receiving red. However, when calculating scores, Homestead ARB, while AF documentation marks the community as yellow, is actually had the lowest student to teach ratio at 17.30 to 1, raising questions as to whether other mistakes made by the AF in their assessment. Forbes Field is again used as the benchmark, upper-tier community with a green mark and receiving a ratio of 12.82 to 1. As percentage of total enrollment, Forbes Field teacher population made-up 7.8% of the district's total enrollment while the community with the lowest student to teach ratio, Homestead AFB, teacher population made up 5.78% of total enrollment. All Florida AF communities had similar ratios, at between 16.07 and 17.30, except for MacDill. MacDill had between 13.96 (without Hillsborough) and 14.66 (with Hillsborough). This is due to, in part, the state's regulation of class size for certain core curriculum. The standard deviation was between 1.39 and 1.29.

The major concern with utilizing this metric is that it speaks only to quantity and not quality, which is a generally theme across all ratios used in the AF's assessment.

Literature Review of Stanford Education Data Archive

Abstract

This literature review aims to discern the relevant academic critiques and thoughts on SEDA data, the model behind it, and other pertinent concerns that may be of interest to Florida and the Department of the Air Force. The exploration of the way this data has been utilized and interpreted is vital to accurately assess the results of the Department of the Air Force's Support of Military Families Education Study. With all data and statistical analysis, there are always a bevy of assumptions made and this review hopes to shed light on some of these. This review will cover both the strengths and weaknesses of SEDA and will delineate a well-rounded picture that presents a hopefully objective picture of the data, and consequently, aid and support the Air Force's study more generally.

Executive Summary

1. SEDA utilizes advanced statistical procedures that attempt to minimize score linking error that have been known for decades; however, the underlying assumptions regarding state test and NAEP score linking remain unchanged. Thus, questions remain regarding the soundness and validity of interpretation about SEDA estimates.
2. As Florida state testing and the NAEP may have different motivations, constructs, and repercussions, which raise validity concerns as to the outcomes used by the AF.
3. The learning rate uses linear prediction to fill in the gaps in years where the NAEP is not taken (it is only taken in grades 4 and 8), but linear interpolation likely has issues. Learning growth is almost certainly not linear, e.g., the learning rate might grow at a faster rate between 3rd and 4th grade than it does between 7th and 8th grades. Since the learning rate accounts for 30% of the overall scoring for the Air Force's study, this is a critical point of contention that deserves further consideration.

Literature Review

The Stanford Education Data Archive (SEDA) was used as the underlying data for the learning rate criterion, which accounted for half of the overall scoring total for Academic Performance and 30% of the overall total that was used to make the relative comparisons amongst AF communities. With the Academic Performance category being weighted so heavily, it places a necessity on all stakeholders to make sure the interpretation and underpinnings of the underlying data are valid and useful for strategic basing decisions.

SEDA and its corollaries are reviewed separately here to make note of the variegated criticisms, thoughts, and opinions about how the SEDA data can and should be interpreted. In this review

there will be cases made for the strengths of SEDA that corroborates its statistical power and validity as well as provide potential weaknesses and limitations of the underlying assumptions about test score linking and the statistical procedures used to create SEDA.

Strengths of SEDA and its Methodology

To develop SEDA, researchers used innovative and advanced statistical procedures to link the state test scores to a common scale, that being the National Assessment of Educational Progress (NAEP). While many assumptions—some cause greater concern than others—underpin the statistical processes at play, it remains integral to highlight some of the principal strengths of the work done by the team at Stanford. Both the SEDA developers as well as outside academics and practitioners will be cited in the subsequent sections.

Methodology of SEDA

SEDA data aimed to provide a rich data set for educational researchers and the public to draw attention to the wide variation across U.S. public schools in student achievement and achievement growth (Fahle & Reardon, 2018). SEDA compiles local achievement information from nearly all students in U.S. public schools and homogenizes these scores into a standardized scale. In doing so, SEDA allows detailed comparisons of school district mathematics and English language arts (ELA) test score *means* and *standard deviations* for students in third through eighth grades. Also, by comparing these means over time from the 2008–2009 to 2014–2015 school years, the SEDA data allow for growth comparison from third to eighth grade in the two disciplines, mathematics and ELA.

The SEDA data utilized advanced statistical procedures that aspired to eliminate—or at least alleviate—prior issues with linking scores on disparate tests. Reardon et al. (2016), clearly outlined their process for linking state scores to the NAEP. The process entails using a heteroskedastic ordered probit model (HETOP) which attempts to estimate the parameters (mean and standard deviation) of the underlying distributions using maximum likelihood estimation. *Heteroskedastic probit* models fit regression models of *ordered* outcomes—such as test proficiency categories—while allowing for *heteroskedasticity (non-constant variance between groups)* in the latent (unobserved) variable. From this, cut scores were determined. Once the cut scores were set, the state-level score distributions were linked to the NAEP scale and then standardized to allow comparison amongst states.

To place the cut scores on a common scale across states, grades, and years, SEDA utilized data from the NAEP. NAEP data provide estimates of 4th and 8th grade test score means and standard deviations for each state on a common scale, as well as their associated standard errors. Because NAEP is administered only in 4th and 8th grades and in odd-numbered years, they had to interpolate and extrapolate linearly (essentially fill in the gaps for 3rd, 5th, 6th, and 7th grades since no data is available) to obtain estimates of these parameters and years, i.e., 2010, 2012, 2014, 2016, and 2018 (Fahle, Chavez, Kalogrides, Shear, Reardon, & Ho, 2021).

Moreover, the SEDA developers describe an approach that allows researchers the opportunity to comprehend more complete information about continuous test score distributions with mean and standard deviation parameters as opposed to solely the raw counts of students in the coarsened proficiency categories. The estimates of these group means and standard deviations can be used to estimate "intraclass correlations (ICCs)," between-group achievement gaps, and other insightful measures such as learning rate (Fahle, Chavez, Kalogrides, Shear, Reardon, & Ho, 2021).

Further, it is essential to focus on learning growth rate, which lies at the heart of the AF's study and is the predominant reason for this review. SEDA used cohort growth (change-in-average growth) to determine learning rates rather than longitudinal growth (average gain score growth) since the latter would require data observations at the individual student level on a year-to-year basis, which is almost impossible to obtain with state test scores. In a perfect world, SEDA estimates could be used as a perfect proxy for longitudinal growth estimates—which most agree are preferred—of student-level growth in cases where the exact same students exist in the same unit (school or district) between two time periods. In practice, this rarely occurs due to student mobility and changes in the makeup of a cohort.

STRENGTHS OF SEDA

Purportedly, SEDA provides detailed comparisons of school district mathematics and English language arts (ELA) test score means and standard deviations for students in third through eighth grades across U.S. public school districts. In addition, by comparing these means from 2008-2015, the SEDA data allow for learning growth comparisons from third to eighth grade in mathematics and language arts skills across school districts (Kuhfeld, Domina, Hanselman; 2019). These growth figures were ultimately used to determine the learning rate used in the AF's education assessment.

Additionally, in a validation study by the SEDA developers, the cohort growth rates—which are explained in the methodology section—correlated highly with longitudinal growth rates, i.e., year to year changes in test scores at the individual level. (Reardon, Papay, Kilbride, Strunk, Cowen, An, & Donohue, 2019). Longitudinal growth rate is the more reliable measure due to its utilization of data with greater levels of fidelity. In consequence, this study concludes that researchers can reliably use cohort growth rates as a proxy for longitudinal growth rates, but there are scenarios where there should be some circumspection (Reardon et al., 2019).

Reardon et al. found that the HETOP model produces unbiased estimates of group means and standard deviations that help inform the cut scores and allow for eventual NAEP linking. The exception was when group sample sizes were small. However, the authors claimed the statistical error that arose from a small sample size could be reduced by using a "partially heteroskedastic" model.

Furthermore, through simulations and real data analyses, Reardon et al. demonstrated that accurate estimation of means and standard deviations of test score distributions for multiple groups (states, districts, schools, etc.) is possible under a wide range of scenarios, with modest loss of efficiency, particularly when sample sizes are larger than 50 and when the cut scores are not highly skewed.

Within the same state, trends on NAEP moved in the same direction as trends on state tests. "States with positive trends between 2005 and 2009 on their own tests tended to show positive trends on NAEP" (Chudowsky, N., Chudowsky V., 2010). However, the trends tended to be greater in magnitude compared to the NAEP. Hence, there does appear to be a positive correlation between state test scores and NAEP, but the mean scores from state tests are outpacing the NAEP score growth.

D. Bolt (2020) stated the results of several validation studies were positive. The predictive accuracy seen both in the Trial Urban District Assessment (TUDA) and Measures of Academic Progress (MAP) analyses, as well as applications to studies of growth and with respect to prediction in non-NAEP grades and years, suggests broad applicability. The aforementioned tests, i.e., MAP and TUDA, are national standardized test similar to the NAEP in content and structure, although some have opined and observed that the NAEP results in lower achievement levels than the MAP. TUDA, on the other hand, was a test created to determine the validity of using NAEP as a "gold standard" for understanding academic progress in urban areas. Therefore, if these tests largely corroborate the work SEDA has done, this provides substantiation to its statistical accuracy and power (Bolt, 2020). However, it should be recognized that these tests may not pose the same impediments that state tests do since they are national standardized tests with relatively similar constructs and intentions to the NAEP; meanwhile, state tests are not constructed uniformly across the U.S.—often with significant differences in difficulty, content, scoring, and test administration—which poses issues that will be touched on later in this review.

Weaknesses of SEDA and Other Related Score Linking Issues

The SEDA data is a highly detailed overview of critical data points that succor the efforts of a plethora of interested parties such as academics, researchers, nonprofits, and policymakers. Although the data has received praise, there remains many issues and limitations that need to be recognized prior to using it as a business decision tool. This section outlines both explicit and implicit weaknesses with SEDA and linking state test scores to the NAEP scale. Insights and remarks from both the SEDA developers as well as other academics and researchers are provided.

WEAKNESSES

In "Uncommon Measures Revisited" (2020), Dorans asserts that it may be occasionally feasible to calculate a linkage between two distinct tests, but a multitude of factors can affect the validity of inferences drawn from the linked scores. "These factors include the content, format, and margins of error of the tests; the intended and actual uses of the tests; and the consequences attached to the results of the tests" (Dorans, 2020). When tests differ on any of these factors, many interpretations of even statistically valid analyses should be considered with skepticism (Dorans, 2020).

Other authors speak even more specifically about the issues with state test to NAEP linking. In "Apples to Apples? The Underlying Assumptions of State-NAEP Comparisons," authors Ho and Haertel offer insights to the specific fallacies and shortcomings of linking the scores of these two types of tests. The article proposes that linkages of state assessments to the NAEP scale probably

involve different constructs, different populations, and tests of different reliability and consistency administered to test takers with different mindsets and motivations. These are less than ideal linking conditions. This issue has been extant since the early 1990s. Robert Mislevy (1992) said that linking often fails not because of how the data is collected and analyzed, but rather because of the differing constructs of the tests being linked.

Put simply, state tests are constructed and intended for a different use than the NAEP. State assessments are administered to all students in specific grades while NAEP state level assessments are administered to selected—yet claimed to be representative—samples of 4th, 8th, and 12th graders. State tests directly measure students' knowledge of *state standards*, which can be highly variable between states. On the other hand, NAEP measures the *cumulative knowledge* of students and not necessarily what they have been taught in the current school year (NAEP FAQ Documentation, 2011). State testing, specifically, has very clear standards that are often taught directly to, i.e., teachers teach to the test. Now, because of this, students are largely inculcated with knowledge that states choose to be present on their tests but may not be anywhere to be found on the NAEP. This results in inconsistencies regarding how state learning rates likely cannot validly be mapped to the NAEP scale. If the content of tests from state to state varies—and sometimes significantly so—then that causes a major flaw in the underlying assumptions made in the SEDA data. Ho and Haertel (2007) illustrate the discrepancy between a state that had high level of proficiency when using their state-based test but had low levels of proficiency when mapping it to the NAEP test. This was explained by the authors as a condition of the state-based test being less rigorous than the NAEP, which inflates scores and allows easier achievement of proficiency. Another reason for the score inflation could be due to another feature of the purpose of state testing, which is to make sure disadvantaged students are keeping up with standards. States often attempt to improve the academic achievement of the disadvantaged under the backing of Title I. Therefore, it becomes reasonable to put emphasis at the point where most disadvantaged students score. On the other hand, the NAEP is much more rigid in its demand and is primarily concerned with "lofty, long-term goals" (Ho & Haertel, 2007).

Many other issues exist that are valid concerns regarding the differentiation between state tests and the NAEP that N. Chudowsky and V. Chudowsky (2010) explicate. First, the classroom instruction is often tailored to the state test content, not the NAEP. This content differential can be drastic, so the score linking is posed with a crucial problem in that the two tests simply are incompatible, and one test cannot be accurately used as a proxy for the other. Secondly, many suspect score inflation on state tests because teachers are teaching to the test, which artificially increases scores and has little to say about true learning growth. Finally, there is more at stake for state tests compared to the NAEP. With this increased importance, both teachers and students may put more effort into preparing and focusing for the test, whereas the NAEP is seen as simply a test to get through with little to no repercussions. For Florida specifically, student test scores can be used to decide whether to allow a student to move to the next grade level or even whether a student graduates, to rate schools and states on how well they are doing, and to bonus teachers (Strauss,

2021). Clearly, there are high stakes for state tests that are non-existent when taking the NAEP. All of these points further solidify the distinction between state tests and the NAEP.

Another notable critique of the score linking between state tests and the NAEP is the time period of test taking. To elaborate, because NAEP is administered only in 4th and 8th grades in odd-numbered years, the SEDA developers linearly interpolated to obtain estimates of the parameters for grades (3, 5, 6, and 7) and years (2010, 2012, 2014, 2016, and 2018) in which NAEP was not administered (Fahle, Chavez, Kalogrides, Shear, Reardon, & Ho, 2021). This is a considerable amount of data that used linear prediction to fill in the gaps, so to speak. This methodology assumes a constant rate of learning growth from year to year; when in actuality this is likely untrue and is backed by research that says marginal learning rates tend to diminish over time, i.e., the learning growth rate from grades 3 to 4 will be greater than in 7 to 8 (Kuhfeld, Domina, & Hanselman, 2019). Since this directly connects to the learning rate that the AF uses, it demands scrutiny about how realistic an assumption of linearity of growth rates really is.

Research reported in Reardon, Papay, Kilbride, et al. (2019) shows that estimates of student learning rates are generally unbiased and reliable, **except when student mobility in and out of schools is high**. The mobility factor may pose a problem for schools with large shares of students who are in military families. High mobility is a constant reality for many military families and their children. Moving from school to school is a common occurrence, and this flaw found in the model may be especially relevant for the AF to consider. Additionally, in very small schools and charter schools, the estimated learning rate is biased upwards, as a result of mobility. More specifically, lower-achieving students more routinely leave schools than enter and this overestimates the predictions. As a result, the SEDA developers recommend that users interpret the school level grade slopes with some caution, especially for charter schools, smaller schools, and other schools that are recognized for high student mobility (Reardon et al., 2019).

To delve briefly into the technical statistical mechanics behind the SEDA data, certain scholars have been unsure about the validity of the heteroskedastic ordered probit model since it uses a technique called maximum likelihood estimation (MLE) to achieve its direct estimates. Lockwood (2018) presents a few caveats and/or criticisms of maximum likelihood estimation exist, particularly when it involves coarsened test score data.

> "Firstly, the MLE has relatively restrictive conditions for its existence (Haberman, 1980; McCullagh, 1980). Existence problems begin to arise when there is at least one group with nonzero counts in fewer than three of the K performance-level categories. With K = 3 or 4 typical in applications involving achievement tests, and with many groups (some of which may be small), it is likely that the MLE of the ensemble of true group parameters does not exist in a given data set" (Lockwood, 2018).

Secondly, the MLE can have large estimation errors when sample sizes are small and/or the marginal probability of one or more of the performance-level categories is small (Lockwood, 2018). This corroborates the SEDA team's finding that the HETOP model does not accurately estimate parameters for

small sample sizes. The estimation errors in the group parameters can lead to noisy and biased estimates of functions of those parameters.

Additional issues can arise in some settings from the fact that the direct estimates do not use covariates, which may include group-level data regarding demographic characteristics of group members and/or information about distributional properties of the true parameters across groups (Lockwood, 2018).

This further solidifies that not only are the more general underlying assumptions about score linkage questionable, but scholars also reserve some trepidation about the underlying statistical procedures used to develop SEDA.

Lastly, there are a few potential issues to recognize with the way the learning growth rate was calculated. These issues were elucidated by the team behind SEDA. As was mentioned in the methodology section of this review, SEDA used a cohort growth rate as opposed to longitudinal growth, and the latter is generally preferred due to greater specificity and exactness. Since the SEDA learning rate is tracking group rather than individual data, there can be some discrepancy between the estimates for the two methods. This is particularly true for schools with high student mobility, i.e., students entering and leaving schools at high rates (Reardon, Papay, Kilbride, Strunk, Cowen, An, & Donohue, 2019). Additionally, poor correlations between the cohort growth and longitudinal growth were found for small schools, predominantly ones with 40 or fewer students in a given grade and year (Reardon et al., 2019). Charter schools' correlation results were also brought into questions due to them typically having both high mobility and smaller student populations (Reardon et al., 2019).

Conclusion

SEDA is responsible for a major portion of the AF's Support of Military Families study on education. The learning rate—which comes from the SEDA data—accounts for 30% of the overall total. Therefore, in order to be certain of a fair and equitable outcome of the study, the underlying data must be scrutinized and examined for any fallacies, weaknesses, or errors. While SEDA's underlying statistical procedures seem to hold up for the most part in a purely academic sense, there remains issues about how to interpret the results due to the host of varying and conflicting constructions, purposes, and other related factors between state tests and the NAEP. However, others claim that the NAEP results are not the gold standard that many assume (Ho & Haertel, 2007). So even in a perfect world, if state test scores could be linked to the NAEP without issue, there still may be an issue of how successfully the NAEP test captures student achievement and performance in the first place.

References

Bejar, Isaac I. *Journal of Educational Measurement*, vol. 47, no. 2, 2010, pp. 255–60. *JSTOR*, http://www.jstor.org/stable/20778951 Accessed 20 May 2022.

Chudowsky, N., and Chudowsky V. "Rising Scores on State Tests and NAEP." *Center on Education Policy*. 2010.

Dorans, N.J., Pommerich, M., & Holland, P.W. (Eds.). (in press). "Linking and Aligning Scores and Scales." Springer.

Dorans, N. J. (2020). Uncommon measures revisited (Educational Testing ServiceResearch Reports ETSRR–20-04). Educational Testing Service.

Fahle, Erin M., and Sean F. Reardon. "How Much Do Test Scores Vary Among School Districts? New Estimates Using Population Data, 2009-2015." *Educational Researcher*, vol. 47, no. 4, 2018, pp. 221–34. *JSTOR*, http://www.jstor.org/stable/44971905. Accessed 20 May 2022.

Haberman, S. J. (1980). "Discussion of regression models for ordinal data." *Journal of the Royal Statistical Society Series* B, 42, 136-137.

Ho, A. D., & Haertel, E. H. (2007). "(Over)-interpreting mappings of state performance standards onto the NAEP scale." Paper commissioned by the Council of Chief State School Officers.

Ho, A. D., & Haertel, E. H. (2007). "Apples to Apples? The Underlying Assumptions of State-NAEP Comparisons." Paper commissioned by the Council of Chief State School Officers.

Lockwood, J. R., et al. "Flexible Bayesian Models for Inferences From Coarsened, Group-Level Achievement Data." *Journal of Educational and Behavioral Statistics*, vol. 43, no. 6, 2018, pp. 663–92. *JSTOR*, http://www.jstor.org/stable/45278326. Accessed 20 May 2022.

Reardon, S.F., Papay, J.P., Kilbride, T., Strunk, K.O., Cowen, J., An, L., & Donohue, K. (2019). Can Repeated Aggregate Cross-Sectional Data Be Used to Measure Average Student Learning Rates? A Validation Study of Learning Rate Measures in the Stanford Education Data Archive. (CEPA Working Paper No.19-08). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp19-08

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., & Chavez, B. (2021). Stanford Education Data Archive (Version 4.1). Retrieved from http://purl.stanford.edu/db586ns4974

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2019). Validation methods for aggregate level test scale linking: A case study mapping school district test score distributions to a common scale. Journal of Educational and Behavioral Statistics. Advance online publication. https://doi.org/10.3102/1076998619874089

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions From Coarsened Data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3–45. http://www.jstor.org/stable/26447647

Strauss, V. (2021, September 18). *Analysis | Florida says it's ending year-end, high-stakes standardized testing. here's what it's really doing.* The Washington Post. Retrieved May 25, 2022, from https://www.washingtonpost.com/education/2021/09/18/florida-desantis-standardized-testing-overhaul/