A Critical Review of the Baseline Soldier Physical Readiness Requirements Study

Executive Summary. The Army's Baseline Soldier Physical Readiness Requirements Study (BSPRRS) was a multiyear effort designed to inform evidence-based change in the Army's physical fitness test of record from a gender- and age-specific standard to a gender- and age-neutral standard based on recurring, physically demanding tasks encountered by soldiers. The study determined Warrior Tasks and Battle Drills and Common Soldier Tasks (WTBD/CST); developed a set of four vignettes that simulated those tasks; identified physical fitness test events that were modeled to predict performance on these vignettes; and finally sought to validate that those test events were indeed predictive. The physical fitness test events determined by BSPRRS became the six-event Army Combat Fitness Test (ACFT). While the Army may have been well-intentioned in its desire to develop an evidence-based, operationally predictive, and gender-neutral fitness standard, close examination of the BSPRRS model reveals critical mistakes and the lack of rigorous cross-validation. Using the ACFT and the BSPRRS model to guide decisions about soldiers' physical fitness may contrarily make for a less lethal Army.

The 2015 NDAA requires that gender-neutral occupational standards "accurately predict performance of actual, regular, and recurring duties of a military occupation" and are "applied equitably to measure individual capabilities." While the Army often claims that the ACFT is over 80 percent predictive, this claim is a misrepresentation of the flawed BSPRRS model. In fact, because researchers likely compared ACFT events against only one of the four WTBD/CST simulation vignettes, only two of the six ACFT test events were modeled as significant predictors of WTBD/CST performance. And, in demonstrating predictive accuracy of the ACFT, the Army used only 136 male soldiers and a mere 16 female soldiers, all volunteers with an average age of 24 years, to represent the entire Army. The Army's new goal of collecting two million data points on soldiers taking ACFT without first addressing the errors in the BSPRRS model or validating the predictive performance of the ACFT is simply shooting in the dark.

Furthermore, the ACFT is not applied equitably. In initial trials with over 14 thousand soldiers, sixty-five percent of all women failed the ACFT, primarily because of the leg tuck test event, compared to ten percent of male soldiers. But, according to data from the Army's own study, leg tucks are not predictive at all of actual, regular, and recurring duties. Indeed, using leg tucks as a criterion creates an unfair adverse impact. The University of Iowa Virtual Soldier Center, who peer reviewed BSPRRS, criticized the Army for the lack of female participants in the study stating that due to the "inherently unbalanced study design...determination of which tasks best predict or represent WTBD/CST performance could be influenced towards strategies used predominantly by men."

Moreover, the ACFT may undermine military readiness. In moving from the original Army Physical Fitness Test (APFT) to the ACFT, the Army has made their fitness test of record 20 times easier for young male recruits and 1.3 times easier for young female recruits. At the same time, the test is more difficult for older female soldiers, precisely those who are already disproportionately underrepresented in senior leadership positions.

Finally, one must question whether the ACFT needs to be gender- and age-neutral at all. While occupational standards are required by the 2015 NDAA to be gender-neutral, the Army emphasizes the role of the ACFT in Holistic Health and Fitness to combat obesity and reduce muscular-skeletal injuries. RAND and others have argued that fitness assessments meant "to maintain a culture of military discipline, bearing, and appearance; to keep health care costs to a minimum; to ensure personnel are not likely to be hampered by chronic illness" can and should be gender- and age-specific assessments.

Background. After forty years of testing physical fitness using push-ups, sit-ups, and two-mile runs, in June 2020 the Army instituted a new assessment, called the Army Combat Fitness Test (ACFT), to be the physical fitness test of record.[1] The ACFT consists of six test events: deadlift, standing power throw, push-ups, sprint-drag-carry, leg tucks, and the two-mile run. The Army is committed to the new test and is already spending over $78 million on the specialized equipment needed to administer it.[2] Every soldier regardless of age or gender is required to complete each test event to one of three minimum standards to pass the ACFT. This is a marked change from the previous Army Physical Fitness Test (APFT) that took age and gender into consideration. Instead, the new standards are based on whether a soldier's military occupational specialty involves a "moderate," "significant," or "heavy" amount of physical exertion.

In explaining the change, former Army Chief of Staff General Mark Milley said "This has everything to do with effectiveness in combat—that's why it's gender-neutral; that's why it's age-neutral. Combat is unforgiving. It doesn't matter how old you are. The enemy doesn't care. Before they shoot you, they don't say: 'Hey, are you 25 or are you 45?' They don't do that. They just shoot you. And dead is dead. So we want to make sure that our soldiers are in top physical condition to withstand the rigors of ground combat."[3]

During the first six months that the Army has been using the ACFT on approximately fourteen thousand soldiers, sixty-five percent of all female soldiers and ten percent of male soldiers have failed to pass the ACFT at the minimum standard.[4] Army Secretary Mark Esper had stated earlier "If you can't pass the Army combat-fitness test, then there's probably not a spot for you in the Army." General Milley had a similar sentiment: "If you can't get in shape in 24 months, then maybe you should hit the road."[5]

ACFT failures were largely due to failing to do one leg tuck, the minimum number needed to pass that test event at the lowest, gold-standard level. Soldiers that must pass gray and black standards must complete even more leg tucks. The Army has insisted that "physiologically, there is no reason any healthy Soldier cannot perform a leg tuck, given 3–6 months of dedicated, regular and progressive training." And "achieving the ACFT GOLD standard takes some Soldiers about 3–4 months of focused training: everyone can meet this standard if properly trained and motivated. We are giving them ~24 months." These statements run counter to advice by expert exercise physiologists.

General Milley has praised the new ACFT stating that it has an "80 percent correlation to the physical activity that is expected of soldiers in the execution of ground combat." Other senior Army leaders promoting the new physical fitness test have made similar claims: "The APFT is a relatively poor predictor (~40%) of a Soldier's ability to execute high demand commonly occurring, critical Warrior Tasks and Battle Drills required of all Soldiers," "The six-event ACFT has been scientifically validated through four years of extensive empirical research (R2 ACFT-High Demand CSTs = 80%) and is a better predictor of physical fitness associated with high physical demand common Soldier tasks,and "Army Combat Fitness Test (ACFT) ~80% ability to predict WTBD/CST performance." These statements are simply not true.

The fiscal year 2015 NDAA requires that gender-neutral occupational standards "(1) accurately predict performance of actual, regular, and recurring duties of a military occupation; and (2) are applied equitably to measure individual capabilities."[7] But the ACFT has neither been shown to accurately predict performance on actual and recurring duties of a soldier nor is it applied equitably. In a 2018 RAND study that examined gender-neutral physical standards for ground combat operations, authors explain equitability in context of physical standards stating that "test validity should not differ among relevant subgroups (such as gender and race), and test scores should be unbiased (i.e., two people who receive the same test score should have the same likelihood of success on the job, regardless of subgroup)"[8] The ACFT fails to meet either of these requirements for a valid, unbiased gender-neutral test.

Statistician George E. P. Box once famously said "All models are wrong, some are useful." The purpose of this review is to examine the ways in which the Army's Baseline Soldier Physical Readiness Requirements Study (BSPRRS) model are wrong and whether the model is indeed useful. Primary background material for this analysis includes the BSPRRS final report,[9] the University of Iowa's review of that report,[10] and earlier technical reports by the U.S. Army Public Health Command[11] and U.S. Army Research Institute of Environmental Medicine.[12]

Phase	Year	Location	Purpose
1	2012–2013	—	Systematic literature review performed by U.S. Army Public Health Center
2	2012–2013	Multiple	Determine the physically demanding, commonly occurring and critical Warrior Tasks and Battle Drills and Common Soldier Tasks (WTBD/CST)
3	2013–2014	Fort Carson	Identify physical characteristics associated with each WTBD/CST to develop a Warrior Task Simulation Test (WTST)
4	2014–2015	Fort Riley	Determine which fitness test events best predict WTST performance
5	2015–2017	Fort Benning	Validate whether fitness test events can accurately predict ability to execute WTBD/CST, are safe to perform, legally defensible, and acceptable

Table 1: Goals and primary field locations of the five phases of the BSPRSS study.[6]

Warrior tasks and battle drills. In an effort to develop an updated and evidence-based fitness assessment, the U.S. Army Research Institute for Environmental Medicine (USARIEM) conducted a major Army-wide fitness prediction study called the Physical Demands Study.[13] As part of their analysis USARIEM identified five domains of combat physical fitness: muscular strength, muscular endurance, aerobic endurance, explosive power, and anaerobic endurance. Using these five domains researchers from the U.S. Army Center for Initial Military Training (CIMT) conducted Army-wide focus groups to identify eleven physically demanding Warrior Tasks and Battle Drills and Common Soldier Tasks (WTBD/CST), such as move as a member of a team and drag a casualty to immediate safety. These eleven tasks were later distilled down to five criterion task vignettes. Soldiers performed a 1.6 km loaded walk/run (the pre-fatigue)[14] and then executed a series of four, timed WTBD simulation test vignettes (WTST):

1. *Build a hasty fighting position.* The soldier filled five 5-gallon buckets with sand using an e-tool. After this they carried sixteen 40-lb sandbags, generally one or two at a time, ten meters and stacked them onto a platform.

2. *Move over, under, around and through.* The soldier completed a course that simulated commonly occurring obstacles in urban/forest terrains: moving 10 m in a high crawl; zigzag running for 45 m while jumping over low obstacles, ditches, and tires; traversing a 24' V-shaped balance beam while carrying a squad automatic weapon and an ammo can; lifting a 50-lb rucksack onto the 48" platform, climbing onto and then moving across the platform, and finally lowering themselves and the object to the ground; scaling a 54" wall; moving over a 42" barrier, under an 18" barrier, through a window, through a 24" × 10' tunnel, and over another 42" barrier; with a combined 50 m sprinting between events.

3. *React to man-man contact.* The soldier performed a set of four obstacles to simulate the physical demands associated with hand-to-hand contact such as pushing, pulling, grasping, and throwing. These obstacles included flipping a 107-lb tire over four times; pushing a 163-lb prowler sled 20 m;[15] lifting and throwing five 30-lb sandbags over a 54" wall; and rotating a 55-gallon trashcan filled with 300 lbs of sand two complete turns clockwise and then two complete turns counter-clockwise.

4. *Extract/evacuate a casualty.* Starting from a prone position beside a barrier, the soldier stood and rushed to a second barrier where they took a knee, and then they completed a short crouch run to a "disabled Humvee"[16] (a 4' × 6' plywood platform with a 47" height and a 2" border), where the soldier extricated a 182-lb training dummy and lowered it to the ground. Finally, the soldier dragged the dummy 20 m to safety and then sprinted 65 m.

The total time of these four events (not including the pre-fatigue walk/run) became a baseline for physical fitness readiness. One might question whether composite vignette completion time alone is the best indicator of fitness.[17] It is not time-on-task that is fundamental to a predictive linear

model but the variance of time-on-task among all soldiers. Furthermore, using a composite time instead of individual vignette times will implicitly weight each vignette by their individual standard deviations.[18]

Predicting a soldier's physical performance. It would be impractical to use the WTST as the fitness test of record for every soldier in the Army, so the researchers looked for possible surrogate physical fitness test events that would be predictive of timed performance on the WTST vignettes. Through a systematic review and the focus group/survey responses, the researchers identified 23 possible predictor test events that would measure muscular strength, explosive power, muscular endurance, cardiovascular endurance, and speed and agility. Many of the tasks of the WTST vignettes required specific physical abilities such as balance, flexibility, and coordination (identified as components of physical fitness by the U.S. Army Center for Health Promotion and Preventive Medicine). For example, traversing the 24-ft V-shaped beam would require a fair amount of balance. Standard physical tests to measure balance and flexibility like standing balance test and sit-and-reach are noticeably absent from the 23 test event candidates.

The researchers' goal was now to determine which smaller set of these 23 test events would be most predictive of the total time to execute all four vignettes. The hope was that with such a model a soldier would need only their scores from these proxy fitness test events to predict their score on the WTST were they to actually take it. To do this, researchers used a linear regression model. Linear regression is a simple approach and probably an overly simple one: start with a bunch of data points, draw the best straight line[19] through those points, and then use that line to make predictions on future events. In the WTBD prediction model, each of a soldier's test event scores (two-mile run time, number of push-ups, maximum deadlift weight, etc.) are multiplied by a conversion factor to change that score into an equivalent number of seconds. Then all of these seconds are added together, along with some baseline time, for a total time meant to predict the composite time of the WTST. It all sounds rather clever, and the Army makes the extraordinary claim that using this model has an "80% ability to predict WTBD/CST performance" from just the ACFT scores. But, the Army has never validated this claim, the model has a number of serious flaws, and the Army does not implement the ACFT in a manner consistent with the model.

Linear regression models have several requirements. One is that outcome variables and predictor variables must be linearly correlated. This means that the weighted predictor scores are simply added together to compute a composite score. So, being especially strong or fast on one test event can offset being particularly weak or slow in another one—a fast run time can offset the number of push-ups or leg tucks. And, unlike the ACFT, linear models do not have maximum scores or minimum passing gold standards. Inability to complete one leg tuck is not a failure.

Ideally, the predictor variables in a linear model are largely uncorrelated with themselves. This reduces the number of redundant variables, simplifying and making the model more explainable. Often great care is taken in the design and selection of predictor variables to reduce multicolinearity. Routine exploratory data analysis techniques such as pairwise scatter plots are used to examine data sets for correlation. And statistical techniques, such as principal component analysis, factor analysis, and variance inflation factor comparison, help to quantify interrelations and identify possible latent variables. Finally, predictive linear regression models require that the data are representative of the population and are homoscedastic, and that the residuals are normally distributed.[20] That the Army has been conducting push-ups, sit-ups, and two-mile runs for every soldier over the past forty years should provide researchers ample data with a minimum baseline to compare the test participants.

Data problems. The BSPRRS model had several other compounding data problems. For example, the casualty-extraction-evacuation vignette was performed in as little as 11 seconds and as much as 516 seconds (almost fifty times longer!), with most times around 90 seconds. When a task that typically takes a minute to complete takes over eight minutes, the quality and representativeness of the data should be examined. Because linear regression minimizes the root mean square error, extreme outliers will have a significant impact on the model fit. The researchers assert that "based on current best practice for regression analyses, individual event scores were not

analyzed or adjusted for distribution abnormalities, which is generally considered to be unnecessary with a least squares model (Fox, 2016)." Their statement is simply not true. Chapter 11, "Unusual and Influential Data," of the cited textbook states emphatically in a bold call-out box centered on the page: "Unusual data are problematic in linear models fit by least squares, because they can unduly influence the results of the analysis and because their presence may be a signal that the model fails to capture important characteristics of the data."[21] When developing any statistical test great care must be taken to understand the data and any outliers that are misrepresentative and would unduly skew the model.

The researchers also state "for incomplete records with minimal missing data, researchers used mean/linear extrapolation to complete the record." These extrapolations led to data peculiarities such as a quarter of all female soldiers as being reported as doing a negative number of pull-ups. It is both mathematically and physically impossible to do anything less than zero pull-ups.

Perhaps the most egregious data problem for a study on developing a gender-neutral predictor model is that the data is not representative of all soldiers in the Army. The initial training data consisted of a group of 290 male and 49 female soldiers (mean age of 24 years with a standard deviation of 4.4 years). For comparison, in 2015 the average military officer was roughly 35 years old and the average enlisted member was just over age 27.[22] The underrepresentation of women during the development of the model was so significant that the Army researchers stated: "following an external review by the University of Iowa, Virtual Soldier Research Center, reviewers suggested we bootstrap additional women into the FT Riley sample to provide a more balanced model and determine if women used a different solution set for WTST performance." Bootstrapping is a technique where data is resampled from already counted data. In effect the researchers simply copy/pasted already overly underrepresented women, virtually cloning an extra 92 women from the original 49. Unlike in electronic music, resampling will not create anything fresh. Even worse, the version of the BSPRRS model that the Army touts as having an 80 percent ability to predict WTBD/CST performance was developed using data from a mere 16 women out of 152 total participants.[23]

event	constant	push-ups	sit-ups	run
coefficient	329.6	−7.19	3.61	0.79
mean	∗	62.8	69.3	885
SD	∗	15.2	10.8	91

Table 2: Linear regression coefficients from Fort Riley training data along with the mean and standard deviations for each of the individual test events for both men and women.

Latent variables. The researchers applied the same ideas to develop a predictive model to examine the original APFT (push-ups, sit-ups, and two-mile run). Table 2 shows the linear regression coefficients derived from the training data along with the mean and standard deviations for each of the individual test events for both men and women.

We can interpret the coefficients as conversion factors, changing the number of push-ups, the number of sit-ups, and two-mile run time into their equivalent seconds of WTST composite time. For example, doing one extra push-up is approximately equivalent to subtracting seven seconds from the WTST, and cutting ten seconds off the two-mile test event is about the same as cutting eight seconds off the WTST time.[24] Oddly, in this model, the number of sit-ups counts against a soldier's WTST predictive performance.[25] Doing one extra sit-up is equivalent to adding 3.6 seconds onto the WTST time. That is to say, the more sit-ups a soldier does, the less fit he or she is perceived as being. Why would the model make such a prediction? It would seem that sit-ups would be a useful measure of core strength and the more sit-ups the more fit, right? Yes, but push-ups, two-mile run, and sit-ups combined are also a predictor of gender. Most men can do more push-ups and can run faster than women. But, men and women are exactly the same on sit-ups. Well, almost. When the researchers tested 278 men and 46 women at Fort Riley, the highest number of sit-ups among men was 102, and the highest among women was 105 (even among a much smaller sample size). According to the Army's model, those 105 sit ups effectively added almost 8 minutes onto this soldier's two-mile run time (or subtracted 52 push-ups).

Statistical models often measure latent or hidden variables. Gender can be a latent variable. When people refer to a machine learning algorithm as sexist or racist, often

it's because the algorithm is finding the latent variables of gender or race buried in the data that it scoops up. Likewise, the Army may have inadvertently built a model trained to look for gender. It seems that the researchers may possibly have realized this, but they state that Army senior leaders directed them to continue with a gender-biased model. There is at least one section in the BSPRRS final report where the researchers possibly acknowledged their concern: "With direction from Army senior leaders, all regression analyses were conducted on the complete sample (both men and women). The reasoning was that baseline Warrior Tasks and Battle Drills and Common Soldier Tasks are criterion tasks that apply equally to men and women."

Height and body mass are themselves latent variables of gender. Women in general are shorter and have lower body mass than men. Female soldiers at Fort Riley and Fort Benning were on average five inches shorter and had 37 pounds less body weight. In designing the WTST events, researchers asked participants to evaluate the difficulty of the tasks and stated "women's concerns related to effects of height and body mass. Taller/higher body mass Soldiers did not identify the same problems." The Army operates as squad/buddy teams rather than as individual soldiers. For example, researchers cited virtually all respondents agreed that scaling a two-meter wall in full fighting load was a two-soldier task and that few Soldiers could scale such a wall in full combat load (85 lbs) as an individual task. In order to design a task that assesses an individual soldier instead of a buddy team, the research team used a modified fighting load weight (50 lbs) and a 1.4-m wall.[26] Under this test scenario taller soldiers have an advantage that may not be representative of scaling a two-meter wall as the buddy team.

When researchers at Fort Carson looked for explicit correlations between weight, height, and fitness variables, they discovered that height and weight among female soldiers were highly correlated to all WTST vignette times except the move-over-under-around-through-obstacles vignette.[27] Height and weight were both negatively correlated, meaning that taller and heavier female soldiers performed faster on these vignettes. Researchers found no significant correlation among female soldiers for fitness variables (sit-ups, push-ups, and run), and they found no significant correlation between height, weight, or fitness variables and WTST time among male soldiers.[28] When researchers asked one professor of kinesiology whether height and weight affect performance on a proposed fitness assessment, he wrote back in all caps "THIS IS BASIC HUMAN PHYSIOLOGY AND SO ROBUST AS TO RENDER IT AXIOMATIC."[29]

Predictive models can reflect biases in subtle ways and when trained on underrepresented groups such models may amplify those biases. BSPRRS Phase 4 (Fort Riley) included only 46 women (14.3 percent) and Phase 5 (Fort Benning) included only 16 women (10.5 percent), a lower representation than even the Army. In their review, the University of Iowa Virtual Soldier Research Center criticized the study for the significant under representation of women, finding that due to the "inherently unbalanced study design. . . determination of which tasks best predict or represent WTBD/CST performance could be influenced towards strategies used predominantly by men."[30]

Coefficient of determination. When the Army makes claims like "The APFT is a relatively poor predictor (~40%)" or "Army Combat Fitness Test (ACFT) ~80% ability to predict WTBD/CST performance," they are referring to the coefficient of determination R^2. The coefficient of determination is a standard measure of the variability of data used to build a model, but it is incorrect to state that R^2 is a measure of predictiveness of the model. Simply stated, R^2 is the percentage of variance in the outcome variable that is explainable or accounted for by the predictor variables.[31] It ranges from zero to one, with $R^2 = 0$ (or zero percent) meaning that there is so much variability in the data that it's impossible to determine any trend, and $R^2 = 1$ (or 100 percent) meaning that all of the data already lines up perfectly and exactly along a line. So, $R^2 = 0.8$ means that 80 percent of the variance in outcome can be explained by the variance in the predictor variables. It does not mean predictor variables have an 80 percent ability to predict performance, as the Army has claimed.[32] Variability in data should be expected, especially when grouping different people together regardless of gender and age. Even among top athletes some are sprinters, some are marathon runners, some are gymnasts—everyone is different.

While adding more variables into the model will always increase its R^2, a more complex model is not necessarily a better one. Therefore, a goal should not necessarily be one that raises the R^2 score of a model, a practice

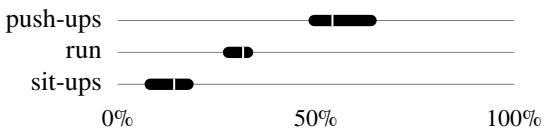

Figure 1: Estimated relative importance of push-ups, two-mile run, sit-ups in predicting WTST performance as a percentage of the R^2. The bars show likely upper and lower bounds of estimation and the notches depict a reasonable point estimate.

often called overfitting. An overfit model may perform wonderfully on the data used to create it, but only on the data used to create it. Once applied to anyone who is not representative of the very narrow training set, the model may be quite wrong. That's precisely what the Army did. The researchers state that the R^2 of the original APFT model is 0.423, saying "the Army Physical Fitness Test (APFT) is a relatively low-to-moderate predictor of WTBD/CSTs performance ($R^2 = 0.423$). . . demonstrating that the APFT is insufficient to ensure Soldiers are capable of performing physically demanding, commonly occurring, and critical Warrior Tasks and Battle Drills and Common Soldier tasks." This statement is simply misleading. The linear regression model in and of itself is the predictor of the WTST performance—not the APFT. Also, the WTST composite time is in and of itself a proxy to the WTBD/CST. This is the same model that counts sit-ups against you and the same model that does not account for variability caused by differences in age and gender.

Furthermore, a high R^2 is not a guarantee that the model is a good representation of the data. Anscombe's quartet is a well-known counter-example that demonstrates that data sampled from very different populations can result in the same predictive linear regression model. See Figure 2. The data in each of the four plots is fundamentally different from one another, yet each one has the same mean, variance, and R^2. When these very different data sets are each used to develop predictive linear models, they all result in exactly the same model, all with the same high $R^2 = 0.666$. But, with exception of the first set, none of the fit lines are good models of the data. The coefficient R^2 itself only describes how much of the total variance in the outcome is accounted for by the predictor variables. It does not tell us how much each predictor variable individually contributes

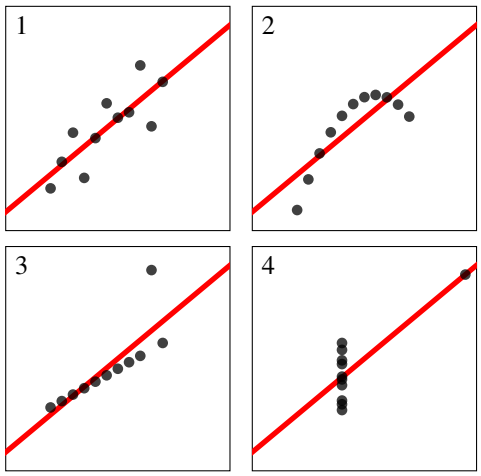

Figure 2: Anscombe's quartet. Different sets of eleven data points resulting in the same linear regression model, all with the same R^2.

to the outcome variance. You should expect that some predictor variables to contribute more than others. So, a natural question to ask is "what is the relative importance of each predictor variable in determining R^2?" One would expect that a predictor variable like two-mile-run time or push-up repetitions to have significant relative importance. Some variables might have little or no relative importance. We can always include any variable as a predictor variable. Preference for peanut butter sandwiches is likely to have zero relative importance as a predictor to the model. Other variables like gender, age, height, and weight all might have high relative importance, but these variables are not explicitly included in the model. Parsing the relative importance of each predictor variable also gets complicated, because each variable may likely be correlated with other variables (except peanut butter sandwich preference,but who knows?). While it is impossible to compute the relative importance of a test event on WTST outcome without having all the data from each soldier that was used to build the linear predictor model, we can estimate it. Figure 1 shows the relative importance of push-ups, run, and sit-ups in determining the performance on the WTST. Each band shows a likely upper and lower estimate and the notch indicates a reasonable estimate. Technical details are provided as an endnote.[33]

Fort Riley. The CIMT researchers' goal at Fort Riley was to determine which smaller set of the 23 test events (pull-ups, vertical jump, dips, etc.) would be most predictive of the total time to execute all four vignettes and build a predictive linear model using those down-selected test events. Data was collected from a group of 290 male and 49 female soldiers (all volunteers) who performed the WTST vignettes along with all 23 fitness test events over the course of several days.[34]

Through a process called stepwise regression, the researchers successively either kept or discarded a test event in an attempt to raise R^2. The use of stepwise regression has a fair amount of criticism, and many argue that it should only be considered as a first step in model selection. It doesn't really take into account expert opinion, and it may enable statistical analysis misuses such as data dredging. Other statistical tests should be used in combination with stepwise regression including F-testing, principal component analysis, or factor analysis.

Through stepwise regression the researchers identified seven test events that raised the R^2 from 0.423, corresponding to the three test events of the original APFT, to 0.737. The events included the sled drag, two-mile run, deadlift, sled push, push-up, kettlebell squat, and power throw.[35] The sled drag, sled push, and power throw measure explosive power. The two-mile run measures speed and cardiovascular endurance. The deadlift and squat measure muscular strength. And the push-up measures muscular endurance. Table 3 shows the regression coefficients.[36]

While the researchers provided a list of seven test events, they neglected to state the relative importance of each test event as a predictor of the outcome. Without the original data, it is impossible to precisely determine the relative importance of each event, but it can be estimated using the coefficients of linear regression and the respective standard deviations.[37] The sled drag is the most significant of the seven predictors of composite WTBD/CST time in the model, and the kettlebell squat is the least. In fact, the sled drag is four times more important than the kettlebell squat as a predictor in the model. The relative importance of the seven test events is shown in Figure 3.

Without having access to the original data set, it is impossible to fully analyze the gender bias that is built into the BSPRRS model. However, we can make a fair estimate of the gender by computing the effect size based

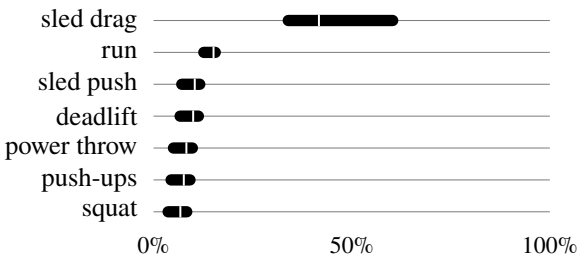

Figure 3: Estimated relative importance of the seven-test-event Fort Riley model as a percentage of R^2.

on differences in the the mean scores of men and women for each test event.[38] One way to interpret the effect size is as the odds that a male soldier chosen at random would outperform a female soldier chosen at random on a given test event. For example, the odds of a male soldier outperforming a female soldier on the two-mile run are 6:1. The odds of a male soldier outperforming a female soldier in push-ups are 10:1, and in sit-ups the odds are an even 1:1. The deadlift is the most gender-biased test event of the 23 candidate events with an odds of 84:1. The gender bias for Fort Riley test events is shown in Figure 4.[39]

The researchers state that "while predictive validity was crucial, it was equally as important to the Army to produce a test that assessed all components of fitness. A multi-component physical assessment was essential to transform physical readiness training and reduce musculoskeletal injuries." Army leadership was concerned that these seven test events "did not represent all components of physical fitness and therefore would not drive a comprehensive change in physical readiness training to increase combat lethality and potentially reduce musculoskeletal injuries." Army leadership was also concerned about the lack of anaerobic endurance and core strength test events. So, the kettlebell squat was removed as a test event and the 300-yard shuttle run and the leg tuck were "forced into the model," although neither are significant predictors of the WTST composite times. Table 4 gives the model coefficients and Figure 5 shows the relative importance of each event in the eight-test-event predictor model.

Under the new model, the sled drag was nine times as important as either the shuttle run or the leg tuck. And both the leg tuck and the shuttle run each contributed to

event	constant	sled drag	run	deadlift	sled push	push-ups	p. throw	squat
coefficient	542.21	10.04	0.41	−0.60	12.29	−1.45	−4.74	−1.45
mean	*	18.3	885	243.9	8.8	62.9	18.3	31.3
SD	*	8.1	91	45.6	2.3	15.2	5.0	13.9

Table 3: Regression coefficients for the Fort Riley seven-test-event predictor model.

event	constant	sled drag	run	deadlift	sled push	push-ups	p. throw	shuttle run	leg tuck
coefficient	436.5	9.67	0.38	−0.80	12.92	−0.98	−4.39	1.67	−1.96
mean	*	18.3	885	243.9	8.8	62.9	18.3	69.2	7.0
SD	*	8.1	91	45.6	2.3	15.2	5.0	5.8	5.0

Table 4: Regression coefficients and importance for Fort Riley eight-test-event predictor model.

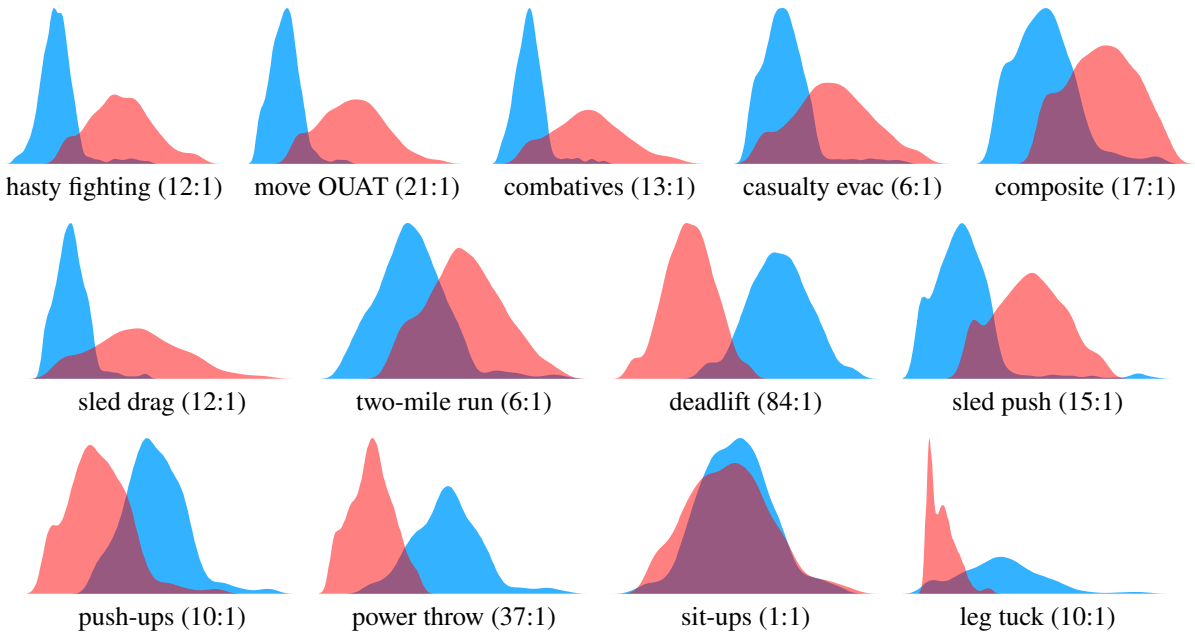

Figure 4: Density plots of Fort Riley fighting-load WTST vignettes and several test events for men and women. The smaller the relative overlap, the greater the gender bias in each event. The odds that a male soldier chosen at random outperforms a female soldier chosen at random are listed in the parentheses.

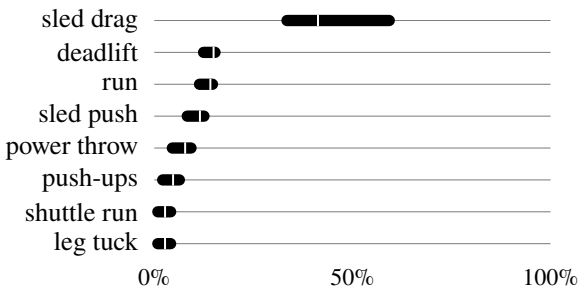

Figure 5: Estimated relative importance of the adjusted eight-test-event Fort Riley model as a percentage of R^2.

less than 5 percent of the models predictive variability.[40]

During testing at Fort Riley, almost fifty percent of female soldiers and five percent of male soldiers were unable to complete even one leg tuck. But, because leg tucks were determined to have such little predictive weight in this model, each complete leg tuck subtracts only two seconds from the predicted WTST composite time. In other words, failure to do one leg tuck is equivalent to adding five seconds onto their two-mile run time or doing two fewer push-ups.

Fort Benning. Model cross-validation is an absolutely essential step in developing a predictive model. Cross-validation tests a model's ability to make predictions given new information about new soldiers and not just the 339 soldiers whose data was already used to make the model. It helps to identify common problems like overfitting and selection bias. And it helps build trust that the model can be generalized across the whole of the Army. To do this, new data (fitness test event scores from a different representative group of soldiers) needs to be collected as input data to the predictor model. The output of the model (their predicted composite WTST times) is then compared with the actual collected data (their actual composite WTST times), and the mean squared error is calculated. Ideally, the mean squared error of the validation data should be close to the mean squared error of the training data.

But, the Army never validated the model. The researchers conducted a follow-up event using 136 male and 16 female soldiers[41] (all volunteers) at Fort Benning with what appears to be the intent of validating the Fort Riley eight-test-event model. Data was collected on WTST completion times and the eight predictive test event scores (the test events that eventually went into building the ACFT). However, instead of using the Fort Benning event to cross-validate the Fort Riley model, the researchers did another "full model regression analysis utilizing the empirical raw scores." It's curious why the researchers, instead of cross-validating the model, decided to create an entirely new model instead. Did the Fort Riley model fail validation using Fort Benning data? Were the test events at Fort Benning so different from test events at Fort Riley as to make comparison of data impossible? By starting over, are the researchers acknowledging a fundamental error in the model design? Table 5 shows the new linear regression coefficients along with data from Fort Benning.[42]

There are several noticeable differences between the Fort Riley and the Fort Benning data. The mean times of the sled drag and the sled push at Fort Riley are 18.3 and 8.8 seconds respectively. At Fort Benning the times were 67 seconds and 33 seconds—about 3.7 times longer. Why was the distance changed for these two test events? Additionally, the mean two-mile run time increased from 885 seconds to 1011 seconds, a change of over a standard deviation. But, the most striking change between Fort Riley and Fort Benning is the model. We can see the difference by simply comparing coefficients. For some unknown reason it appears that rather than using the composite WTST time for all four vignettes, the researchers used only the time for the first, hasty-fighting-position vignette to develop their new predictive linear regression model. To see this, we can apply the model to the mean test event scores to give a mean predicted time of 268 seconds, which agrees closely with the average WTST hasty-fighting-position mean of 262 seconds, and is quite different from the composite time of 606 seconds. For further confirmation we can also compare the standard deviations in the test event scores and the WTST vignette times.[43] The model predicts a standard deviation of about 44 seconds, which agrees moderately well with the average WTST hasty-fighting-position standard deviation of 67 seconds, compared to a 202-second standard deviation across all four vignettes.

In the Fort Benning model, the two-mile run accounts for almost 70 percent of the variance. The leg tuck and the push-up on the other hand account for less than one percent. Furthermore, the coefficient of the leg tuck is

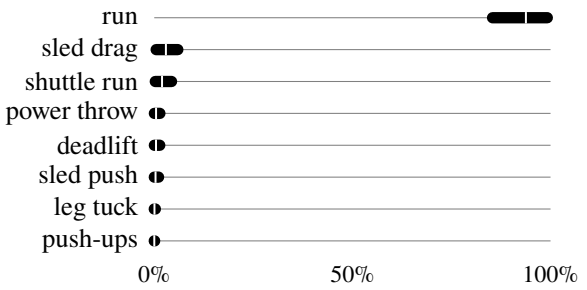

Figure 6: Estimated relative importance of the eight-test-event Fort Benning model as a percentage of R^2.

Figure 7: Estimated relative importance of six-test-event Fort Benning model as a percentage of R^2.

positive, saying that it is anti-correlated with improved WTST scores. The leg tuck and push-up should not be used as predictors in the model (which applies to the hasty-fighting-position vignette).

Following the conclusion of the study, Army senior leaders asked that the sled drag, sled push, and shuttle run be combined into one test event, the sprint-drag-carry (SDC), to reduce administration time, equipment costs, and the total number of test events. The SDC is composed of a 50-m sprint, a 50-m sled drag, a 50-m sideways lateral dash, a 50-m farmers carry, and finally an additional 50-m sprint. There are clear concerns about making such a substitution—it is not the same test event! In an attempt to incorporate this post hoc change into the Fort Benning model, the researchers consolidated the sled drag, sled push, and shuttle run variables into one composite variable based on the standardized values of these three test events.[44] Table 6 shows the new linear regression coefficients along with data from Fort Benning.[45] Figure 6 shows the relative importance of these test events in the model.

The final report does not provide data on the amalgamated SDC test event outside of stating that the data was computed using the z-scores of the sled push, sled drag, and shuttle run. The mean and standard deviations listed in Table 6 for the SDC are likewise derived using z-scores to combine the point estimates from these three events.[46] The relative importance of these test events in the adjusted model are shown in Figure 7. The score is not representative of a real sprint-drag-carry test event. The sprint-drag-carry and two-mile run now dominate the six-test-event model as predictors. The other four events are all less than one percent. Moreover, both the leg tuck and push-up have positive coefficients. Saying that leg tucks (or push-ups) have any real relevance as predictors in the model is a bit like giving credit to an airline passenger for a safe landing, simply because he was included on the flight manifest along with the pilot and other crew members. So, there it is—the model that the Army has claimed as being 80 percent predictive.

The Army Combat Fitness Test. In moving from a predictive model to a physical fitness evaluation framework, there are several things that should be remembered. A predictive linear regression model does not have minimum scores, only a cumulative score. Being faster or stronger in one category can offset being slower or weaker in another. The test events in a linear model are each weighted differently and determined by the relationship of the predictor variables to the outcome variables. The two-mile run and the sprint-drag-carry are the most predictive events, while leg tucks and push-ups are the least. In the Fort Riley model leg tucks have a relative importance of four percent and in the flawed Fort Benning model leg tucks are less than one percent, while the sprint-drag-carry and two-mile run are together well over 70 percent. Women were vastly underrepresented in the model training data. The model that the Army erroneously claims as having an 80 percent ability to predict WTBD/CST for every soldier was developed using a mere 16 women. Moreover, the model was developed using volunteers who are not representative of the Army as a whole. In the Army's attempt to construct a gender neutral-test, they eliminated sit-ups, the one test that was truly gender neutral. Half of all women in the training couldn't do a leg tuck.

The ACFT consists of six timed events, each based on

event	constant	sled drag	run	deadlift	sled push	push-ups	p. throw	shuttle run	leg tuck
coefficient	6.72	0.049	0.28	−0.012	0.016	−0.003	−0.099	−0.27	0.02
mean	∗	67.0	1011	229.5	33.1	52.5	20.6	70.5	9.2
SD	∗	48.3	122	46.2	27.0	14.3	5.7	6.5	5.8

Table 5: Regression coefficients and importance for Fort Benning eight-test-event predictor model.

event	constant	SDC	run	deadlift	push-ups	p. throw	leg tuck
coefficient	12.21	1.20	0.16	−0.015	0.021	−0.082	0.031
mean	∗	*82.4*	1011	229.5	52.5	20.6	9.2
SD	∗	*55.8*	122	46.2	14.3	5.7	5.8

Table 6: Regression coefficients and importance for final Fort Benning six-test-event predictor model.

tasks that a soldier might encounter in training or combat.[47] The Army has claimed that the deadlift replicates a "litter carry or the movement of ammunition and supplies." The power throw replicates "the movement required to assist a buddy over an obstacle or the power required to leap across a ditch." The sprint-drag-carry replicates "moving a casualty to safety moving supplies or moving under fire." The push-up replicates "hand and arm movements required in combatives or repetitive loading of ammunition and supplies." The leg tuck replicates "climbing up and over walls obstacles or exiting disabled vehicles." And the two-mile run replicates "movements to and from contact." Every soldier must pass each test event at a minimum gold, gray, or black standard depending on his or her military occupation. See Table 7. Table 8 shows the percentage of soldiers in Fort Riley trials achieving at gold, gray, or black standard scores for male and female soldiers. [48,49] The bottom row shows the percentage of soldiers reported by the Army during the first six months of ACFT implementation.[50]

During six months of ACFT trials, with approximately fourteen thousand soldiers, 60 percent of all female soldiers and eight percent of all male soldiers were unable to do one leg tuck (the minimum number needed to pass.) See Table 8. Shortly afterwards, due to the incredibly high leg-tuck failure rate, the Army allowed service members to temporarily substitute a two-minute plank in lieu of a leg tuck.[51] The plank was never included as one of the 23 test events examined during the development of the BSPRRS model. Exercise physiologists have noted that a plank and a leg tuck are very different assessments. A plank is an isometric assessment of core fitness. A leg tuck is a dynamic assessment of "grip strength, shoulder adduction and flexion, elbow flexion, and trunk and hip flexion."[52] That the Army is substituting one test event with another test event of different muscle groups underscores the weak and arbitrary reasoning behind the leg tuck in the first place.[53] A plank hardly replicates "climbing up and over walls obstacles or exiting disabled vehicles."[54]

Army Chief of Staff General Mark Milley said of the new ACFT "This fitness test is hard. No one should be under any illusions about it." He's correct, unless he's referring to the roughly 40 percent of soldiers who are male and under twenty five. Under the original APFT an 18-year old male soldier needed to do 42 push-ups to pass. But, under the new ACFT standards that same soldier needs to do only ten.[55] Put another way, an 18-year old under the ACFT standards needs to have the same upper-body muscular endurance as a 50-year-old female soldier and less than a 65-year-old male soldier, who needed to do an additional six push-ups under the old APFT.[56] That 18-year-old soldier also doesn't need to have the same speed and aerobic endurance under the new ACFT as he did under the old APFT. The ACFT two-mile gold standard run time is 21 minutes.[57] That 18-year-old male soldier under the ACFT is allowed an extra minute over what a 65-year-old man needed under the APFT rules. At the same time, the ACFT forces a 40-year-old woman to run

standard	score	deadlift	power throw	two-mile run	push-ups	leg tuck	SDC
maximum	100	340	12.5	13:30	60	20	1:33
black	70	200	8.0	18:00	30	5	2:10
gray	65	180	6.5	19:00	20	3	2:30
gold	60	140	4.5	21:00	10	1	3:00
minimum	0	80	3.3	22:48	0	0	3:35

Table 7: ACFT standards and scoring. Deadlifts are counted in pounds, power throw in meters, two-mile run and SDC in minutes, and push-ups and leg tuck in repetitions.

standard	deadlift	power throw	two-mile run	push-ups	leg tuck	SDC
maximum	*0* | *0*	*96* | *23*	*23* | *1*	*69* | *8*	*2* | *0*	*
black	*88* | *3*	*100* | *80*	*96* | *81*	*100* | *80*	*74* | *3*	*
gray	*96* | *11*	*100* | *94*	*98* | *97*	*100* | *100*	*84* | *19*	*
gold	*100* | *66*	*100* | *100*	*100* | *100*	*100* | *100*	*92* | *56*	*
FY19	95 | 92	96 | 94	96 | 84	96 | 94	90 | 17	82 | 82
FY20-Q1	98 | 94	99 | 96	97 | 90	99 | 96	92 | 36	98 | 88
FY20-Q2	100 | 99	100 | 97	98 | 91	100 | 97	93 | 46	99 | 94

Table 8: Percentage of soldiers in Fort Riley trials achieving at gold, gray, or black standard scores, split by male (left) and female soldiers (right). The bottom row is the percentage of soldiers reported by the Army.

almost two minutes faster. The original APFT minimum standards were constructed using normative measures. Thirty percent of 40-year-old female soldiers ran in over 21 minutes.[58] Even when training regularly, runners run slower with age.[59] And many older soldiers have chronic injuries resulting from years of deployments and carrying heavy gear.[60] If the Army argues that every soldier must have the same levels of minimal fitness regardless of age, and that indeed, with a few months of training, any soldier can achieve gold standard, why does the Army set the maximum age to enlist for active duty at 34 years?

In April 2020 the Army tested a battery of basic trainees at Fort Sill, Oklahoma using both the APFT and the ACFT. They found men had a 60 percent fail rate on the APFT but only a 3 percent fail rate on the ACFT. Women also had 60 percent fail rate on the APFT but a 47 percent fail rate on the APFT.[61] The Army has effectively made their fitness test of record 20 times easier for male recruits and 1.3 times easier for female recruits. The fail rates on the APFT were largely due to the two-mile-run, and the fail rates on the ACFT were largely due to the leg tuck. In going from the APFT to the ACFT, the Army has made the two-mile-run much easier, especially for young male soldiers. At the same time, they added leg tucks that predominately adversely affect female soldiers.

The 2015 NDAA requires that gender-neutral occupational standards first "accurately predict performance of actual, regular, and recurring duties of a military occupation," and second "are applied equitably to measure individual capabilities." RAND has offered the following two criteria for assessing where physical fitness tests are equitably applied: "Test validity should not differ among relevant subgroups (such as gender and race), and test scores should be unbiased (i.e., two people who receive the same test score should have the same likelihood of success on the job, regardless of subgroup)." The ACFT fails to meet either of these requirements for a valid, unbiased gender-neutral test.

Adverse impact, often referred to as unintentional discrimination, is any employment policy or practice that disproportionately and adversely affects one group of peo-

ple of a protected class (such as sex or age) more than another, even though the rules are formally neutral. U.S. labor law prohibits employers from using tests or selection procedures that are not "job-related for the position in question and consistent with business necessity" and that demonstrate adverse impact. While there is no single specific threshold or test to prove discrimination on the basis of adverse impact, a typical metric is the four-fifths rule. The rule states that if the selection rate for a certain group is less than 80 percent of that of the group with the highest selection rate, there may be adverse impact on that group. The four-fifths rule does not apply to selection of military personnel, but similar principles hold given the military's commitment to equal opportunity. The average gold-standard pass rates of the ACFT measured over six months of initial testing on over 14 thousand soldiers was 32 percent for women and 89 percent for men. The Army did not provide further disaggregation by age, nor did the Army provide a breakout using gray or black standards, which are higher rates for certain career fields.[62] The impact ratio (32 divided by 89) is 36 percent, which is much lower than the 80 percent threshold needed to demonstrate adverse impact.

In assessing test fairness RAND researchers argue that "adverse impact alone does not indicate that a test is unfair to the group affected. A test could show adverse impact for women, but it could still be a fair and accurate predictor of their ability to do the job." Instead, they cite predictive bias as a second consideration of fairness. Predictive-validity bias refers to a test's accuracy in predicting the performance of a group of soldiers. A test would be considered "unbiased" if it predicted performance equally well for all groups of soldiers. If a test is a better predictor of one group than another, then the test is considered biased against the group with lower predictive validity. It is quite clear that the leg tuck test event is biased against women. Leg tucks, which have a 92-to-40 male-to-female gold standard pass rate, were shown to have no impact on predictive performance in the Fort Benning model. Even in the Fort Riley model leg tucks account for only four percent of predictive performance. In other words, according to the predictive model, each leg tuck is equivalent to roughly two push-ups, a two-pound increase in deadlift, or five seconds of the two-mile run. Yet, in going from gold to gray standard on the ACFT, the Army has set each leg tuck equivalent to roughly ten push-ups, a twenty-pound increase in deadlift, or sixty seconds of the two-mile run. The best strategy for a soldier to maximize his or her ACFT score, is by doing more leg tucks, a strategy that is most easily accomplished by male soldiers, yet one that has been shown to have little if any impact on the predicted performance on the WTST.

For years, the military services have had different standards for general fitness and for occupational requirements. RAND researchers argue: "A fitness standard developed with the goal of improving overall health may determine that minimally acceptable fitness levels could be higher for younger, male personnel. In contrast, fitness standards developed with the goal of ensuring physical readiness to perform occupationally relevant, physically demanding tasks may use one standard for all personnel expected to perform those physically demanding tasks."[63] In another study RAND researchers argue that military services traditionally have set two types of physical standards. General fitness standards, used to promote overall health status and physical fitness regardless of occupation, need not be gender neutral. The researchers state that "it could be reasonable to continue to use gender-specific standards for physical fitness requirements for enlistment and continuation in service, while still requiring that occupation-specific physical standards be gender neutral, if the goals of the two standards are different." They explain that the goal of occupation-specific standards is "to ensure that people are capable of performing a specific set of tasks required of everyone or considered critical to performance on the job," regardless of gender or age. Fitness standards, on the other hand, could be in place "to maintain a culture of military discipline, bearing, and appearance; to keep health care costs to a minimum; to ensure personnel are not likely to be hampered by chronic illness; and to ensure that the personnel hired reflect the portion of the U.S. population who are at the peak of their health." Specifically, they state "All of these goals can be achieved using screening tools that evaluate someone's overall medical health and fitness. However, in general, medical health and fitness measures used for these purposes are gender-normed and age-normed."[64] The Army itself has argued that the the ACFT is part of its broader Holistic Health and Fitness Program, whose stated goals are to improve individual Soldier readiness, transform the Army culture of fitness,

reduce preventable injuries and attrition, enhance mental toughness and stamina, and contribute to increased unit readiness.[65] From the Army's own framework, the ACFT is best categorized as a general fitness assessment and as such ought to be gender- and age-normed.

Recommendations. It is clear under any scrutiny that the BSPRSS predictive model is wrong. Recounting just one example—the model predicts a mean performance of 260 seconds on the WTST using ACFT test events, when the true mean performance time was over 600 seconds. That neither Army researchers nor University of Iowa reviewers caught such a gross error indicates that the model was not adequately scrutinized. The 2015 NDAA requires that occupational standards accurately predict a soldier's performance on actual duties and that they are applied equitably. Even when corrected for gross errors, it is unlikely from the Army's own study that leg tucks have any substantive predictive measure in the model. Returning to George Box's aphorism about the wrongness and usefulness of models, the following are recommendations for Army researchers and leadership on making the BSPRSS model less wrong and more useful:[66]

- Pause implementation of the ACFT as fitness test of record until the errors and concerns can be addressed. Reexamine the BSPRRS model and the training data to correct all gross errors and account for anomalies.
- Establish ACFT success criteria based on success criteria on WTST or other real-world tasks. Because success criteria were never identified on WTST vignettes, there is no way set non-arbitrary ACFT standards. For example, what is the maximum acceptable number of seconds allowed to extract and evacuate a casualty or to move over an obstacle?
- Be truthful about the model performance and limitations. Stating that the ACFT is over 80 percent predictive is bullshit.[67]
- Openly publish the model and data in machine-readable format to increase transparency and follow best practices in evidence-based policymaking.[68] Use exploratory visualizations to help identify patterns and anomalies in the data and to make the results accessible and compelling to broad audiences.
- Perform factor and senstitivity analysis of the ACFT events and WTST vignettes along with other variables such as height, weight, body mass, and gender to determine the relative importance of each variable and identify hidden biases.
- Cross-validate the BSPRRS model using soldiers who are representative of all ages, genders, and military occupations. Better yet, use a representative sample of soldiers in developing a new model.
- Think critically about model design and data collection. Continue working with external experts such as the University of Iowa Virtual Soldier Center in developing a scientifically rigorous standard. Include a diverse range of experts early in the design process to avoid group think. Red team the model to identify shortcomings.
- Clearly delineate the function of the ACFT as either an operationally predictive, gender-neutral standard or as a component of the Holistic Health and Fitness Program meant to address injuries and general fitness.

Notes.

[1] Ryan D. McCarty, *Army Directive 2020-06: "Army Combat Fitness Test"*, June 2020

[2] https://www.army.mil/acft/#faq-section-5

[3] Jeff Schogol, "Army Chief Tells Soldiers To Get Fit Or Get Out," *Task & Purpose*, October 2018,

[4] Gold-standard pass rates on the ACFT (from Army memo dated 09 July 2020 in response to Senator Gillibrand's request for information):

	FY19	Oct	Nov	Dec	Jan	Feb	Average
Female	21	37	29	33	34	40	32
Male	81	91	89	91	90	90	89

[5] Matthew Cox, "Army Chief: Pass New Combat Fitness Test or 'Hit the Road'," *Military.com*, October 2018,

[6] The BSPRSS final report combines the first three phases into one phase and instead references Phases I, II, and III.

[7] *National Defense Authorization Act for Fiscal Year 2015*, December 2014

[8] Chaitra M. Hardison, Susan D. Hosek, and Anna Rosefsky Saavedra, *Establishing Gender-Neutral Physical Standards for Ground Combat Occupations: Volume 2. A Review of the Military Services' Methods* (The RAND Corporation, 2018)

[9] Whitfield B. East, MAJ David DeGroot, and Stephanie Muraca–Grabowski, *Technical Report: T19.041-13.1 Baseline Soldier Physical Readiness Requirements Study* (Fort Eustis, VA: Research / Analysis Division, U.S. Army Center for Initial Entry Training, November 2019)

[10] Karim A. Malek et al., *Review Report: Baseline Soldier Physical Readiness Requirements Study* (The University of Iowa, April 2020)

[11] Bruce H Jones et al., *Development of a New Army Standardized Physical Readiness Test: January 2012 through December 2013* (2015)

[12] Jan E. Redmond et al., *Development of the Occupational Physical Assessment for Combat Arms Soldires (USARIEM Technical Report T16-2)*, Military Performance Division, U.S. Army Research Institute of Environmental Medicine, October 2015; Jan E. Redmond et al., *Development of a Physical Employment Testing Battery for Artillery Soldiers: 13B Cannon Crewman and 13F Fire Support Specialalist (USARIEM Technical Report T16-9)*, Military Performance Division, U.S. Army Research Institute of Environmental Medicine, December 2015; Jan E. Redmond et al., *Development of a Physical Employment Testing Battery for Infantry Soldiers: 11B Infantryman and 11C Infantryman - Indirect Fire (USARIEM Technical Report T16-10)*, Military Performance Division, U.S. Army Research Institute of Environmental Medicine, December 2015

[13] Jones et al., *Development of a New Army Standardized Physical Readiness Test: January 2012 through December 2013*

[14] This was used to simulate a 10 km road march intended to replicate the physical pre-fatigue created by moving over uneven terrain to the objective.

[15] This obstacle was changed from a 20-m power drag of a 163-lb weighted evacuation sled.

[16] The researchers originally used an actual Humvee for the casualty-extraction-evacuation vignette. The time to get the combatives dummy back into the Humvee between vignettes was so long as to be impractical that they changed the casualty extraction to a wooden bench seat and later to a flat table.

[17] The University of Iowa reviewers did question this metric.

[18] The composite standard deviation will be between the sum of the standard deviations (L^1) and the square root of the sum of the squares of the standard deviations (L^2) depending on the degree of correlation between vignette times. For $L^{1.3}$ the relative importance of each vignette is hasty fighting (37%), move OUAT (14%), combatives (27%), and casualty evacuation (22%).

[19] For multiple regression, the line is a hyperplane.

[20] A least squares regression solution is the maximum likelihood estimator of normally distributed data.

[21] John Fox, *Applied regression analysis and generalized linear models* (Sage Publications, 2015), p. 270

[22] Pew Research Center, *Facts About The U.S. Military's Changing Demographics*, 2020

[23] Women appear to be overlooked so much in the BSPRR study that while there were several analysis and comparison tables in the final report dedicated only to men, there were no equivalent tables for women.

[24] Expected value is a linear operator. By taking the dot product of the coefficients and the means and then adding the constant term, we get a total of 831 seconds, which agrees closely with the average composite WTST mean time of 842 seconds. Without knowing more about the correlation of the variables, we can't transform the variances, but we can make an estimate. If the variables were linearly independent, their covariances would be zero and the standard deviation is a weighted L^2-norm of standard deviations. If the variables were highly correlated, then the standard deviation would be a weighted L^1-norm of the standard deviation of the independent variables. By taking the dot product of the absolute values of the coefficients and the standard deviations, we get a total of 220 seconds. Computing the weighted L^2-norm (square root of the dot product of these terms squared) we have 137 seconds. The average composite WTST standard deviation is 234 seconds. The difference between the predicted and the average lies in the explained variation.

[25] One might question whether there is a typo or a sign error in the table. But, after reconstructing the WTST mean time and standard deviation from those of the predictor variable, it is clear that the numbers in the table are correct.

[26] I am 70 inches tall (the average height of a male soldiers participating in the test events). To me a 1.4- m wall is the good height to get proper leverage. If I were five inches shorter (the average height of a female soldier), scaling such a wall would be considerably more challenging.

[27] Reseachers ranked correlation with Pearson correlation values. Height and weight typically scored $R < -0.5$ for female soldiers and $R \approx -0.2$ for male soldiers.

[28] Jones et al., *Development of a New Army Standardized Physical Readiness Test: January 2012 through December 2013*, pp. 13–15

[29] Jones et al., p. E–16

[30] A. Malek et al., *Review Report: Baseline Soldier Physical Readiness Requirements Study*, p. 10

[31] There are several equivalent definitions for R^2 including as the percentage of the explained variance and as the square of the Pearson correlation coefficient between the observed outcome variables and the predicted outcome variables.

[32] The BSPRRS final report includes p-values when it mentions R^2. A p-value is the likelihood that the outcome data and the predictor data are uncorrelated, i.e., the likelihood that R^2 equals zero or equivalently that all the regression coefficients are zero. In the null hypothesis, the R^2 is distributed as a beta distribution $\text{Beta}((k-1)/2, (n-k)/2)$, where k is the number of coefficients including the constant term and n is the sample size. When $n = 300$ and $k = 6$, we find that even a p-value of less than 0.001 whenever R^2 is greater than 0.07. In other words, the p-value is not a meaningful statistic.

[33] $R^2 = \text{Var}\left[\sum_{i=1}^n a_i X_i\right] / \text{Var}[Y]$ is the variance in the outcome variables accounted for by the predictor variables divided by the total variance in outcome. Let's examine

$$S = \text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i,j=1}^{n} a_i a_j \text{Cov}\left[X_i, X_j\right].$$

If the variables X_i are all uncorrelated, then $\text{Cov}[X_i, X_i] = \text{Var}[X_i]$ and $\text{Cov}\left[X_i, X_j\right] = 0$ for $i \neq j$. It follows that

$$S = \sum_{i=1}^{n} a_i^2 \text{Var}[X_i] = \sum_{i=1}^{n} a_i^2 \sigma_i^2,$$

and the relative importance is

$$\text{RI}^2 = \frac{(a_i \sigma_i)^2}{\sum_{i=1}^n (a_i \sigma_i)^2}.$$

Now, consider the case when the variables X_i were all completely correlated. That is, Pearson's correlation coefficient $\rho_{ij} = 1$ where the

$\rho_{ij} = \mathrm{Cov}\big[X_i, X_j\big]/\sigma_i\sigma_j$. It follows that

$$S = \mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \left(\sum_{i=1}^{n} a_i \sigma_i\right)^2.$$

A simple way to partition S is by taking: $S_i = a_i\sigma_i \sum_{i=1}^{n} a_i\sigma_i$. In this case, the relative importance is

$$\mathrm{RI}^1 = \frac{|a_i\sigma_i|}{\sum_{i=1}^{n}|a_i\sigma_i|}.$$

If all the terms were completely dependent, linear regression would also be quite ill-conditioned. Realistically, variables will have some degree of correlation, but not be completely correlated. Heuristically, a suitable estimate for the relative importance is taking

$$\mathrm{RI}^p = \frac{|a_i\sigma_i|^p}{\sum_{i=1}^{n}|a_i\sigma_i|^p}$$

for some $1 \le p \le 2$. Empirically, we can try several values of p to find which one gets S closest to $R^2\mathrm{Var}[Y]$. We find that $p = 1.3$ does a pretty good job, meaning that the variables tend to be a little more linearly dependent than linearly independent. The ends of the segments in Figures 1 and 3–7 are determined by computing RI^1 and RI^2 and the notches are determined by computing $\mathrm{RI}^{1.3}$. (It may happen that terms are anticorrelated, in which case covariance terms are negative, resulting in compounding effects of the variables. Without the actual data or at very least a covariance matrix, it is difficult or impossible to know such effects. Covariance matrices are really not difficult to print and should be included!)

[34] The WTST vignettes were conducted three times: a practice trial in combat uniform only with 285 men and 46 women, a trial in modified fighting load without pre-fatigue (a 1.6 km loaded walk/run with a 55–65-lb load) with 277 men and 44 women, and modified fighting load with pre-fatigue with 256 men and 35 women. The researchers used the average of the two fighting-load trial times as the dependent variable for the linear regression. Presumably, participants dropped out of trials and it is unclear how the researchers averaged the missing data or performed regression using missing data. Typically, missing data will need to be removed from the data set. The researchers noted the Pearson correlation factor between the two modified fighting load trials was $R = 0.833$. It seems problematic that $R^2 = 0.639$ between the tests is smaller than the $R^2 = 0.737$ of test events. This may indicate overfitting. See page 28 of the BSPRRS final report.

[35] The BSPRRS final report is inconsistent about whether stepwise regression yielded seven or eight test events. Page 28 of the report states: "The stepwise linear regression model identified eight variables that accounted for a relatively high percentage of explained variance for WTST performance; $R^2 = 0.737$; $p < 0.05$ (see Table 14). The eight variables were: sled drag, power throw, two-mile run, deadlift, sled push, leg tuck, kettlebell squat, and push-up." However, Table 14 only includes seven variables and does not include leg tucks. Table 15, which swaps the shuttle run in for the kettlebell squat and includes the leg tuck, states: "Adjustments made to ensure proper physiologic balance to include anaerobic endurance, core strength training." The shuttle run was added to measure anaerobic endurance, and presumably leg tucks were added to measure core strength. Computing the predicted WTST times using the averaged test event scores and coefficients in Tables 14 and 15 of the BSPRRS final report yields 829.5 and 831, respectively. The coefficients are listed with three digits of precision after the decimal, leading to a potential maximum round off error of 1.6. So, the tables are in agreement when Table 14 does not include leg tucks. Furthermore, the level of relative importance of the variables in the seven-test-event predictor model and the eight-event-predictor model are roughly the same. The least important variables in the eight-variable model are the leg tuck and shuttle run. Furthermore, the University of Iowa BSPRSS review report states that the Army initially identified seven test events and later added leg tucks. So, there is strong evidence to conclude that leg tucks were not originally derived through stepwise regression but were later "forced into the model" along with the shuttle run.

[36] By taking the dot product of the coefficients and the means added to the constant term, we get a total of 830 seconds, which agrees closely with the average composite WTST time of 842 seconds. Taking the dot product of the absolute values of the coefficients and the standard deviations, we get a total of 240 seconds, which agrees closely with the average composite WTST standard deviation of 234 seconds.

[37] The relative importance is the percentage contribution of each predictor variable to R^2. Even if we don't know the covariance of the variables, we can estimate bounds by using the weighted L^1 or L^2 norms of the standard deviations.

[38] In making effect size readily understood by non-statisticians, researchers McGraw and Wong (1992) have suggested using a Common Language Effect Size (CLES), which gives the probability that a score sampled at random from one distribution would be greater than a score sampled at random from another distribution. This probability is the standard normal cumulative distribution function evaluated at $d/\sqrt{2}$ where d is Cohen's standardized mean difference d. Cohen's d is the difference of the two means divided by the pooled standard deviation: $d = (\overline{x}_1 - \overline{x}_2)/s$ where $s = \sqrt{(n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2)}$. This makes a strong assumption on normality of the data, the same assumptions that the BSPRRS researchers make. An alternative approach that doesn't assume normality is by sampling from the distribution. The BSPRRS final report lists scores at several percentiles for each test event: 0, 5, 10, 25, 50, 75, 90, 95, 100. We can use a piecewise cubic Hermite interpolant (or a piecewise linear inerpolant) through the points to reconstruct the distributions and then use the Monte Carlo method to draw samples of female and male soldiers for comparison. To make CLES even more readily understood, we can reinterpret it as an odds $p : 1 - p$, and then normalizing the ratio $p/(1 - p) : 1$.

[39] The density plots are derived by Monte Carlo sampling ($n = 1000$) the distribution reconstructed using a piecewise cubic Hermite interpolant (pchip) of the percentiles listed in the BSPRRS final report. The curve is smoothed using a Gaussian kernel.

[40] By taking the dot product of the coefficients and the means added to the constant term, we get a total of 831 seconds, which agrees closely with the average composite WTST time of 841 seconds. Taking the dot product of the absolute values of the coefficients and the standard deviations, we get a total of 235 seconds, which agrees closely with the average composite WTST standard deviation of 234 seconds.

[41] It's concerning that the Army used only 16 women to represent fitness standards that impact 74,000 women. Women accounted for less than 11 percent of the study participants, less than even the 15 percent composition of the Army.

[42] The data from Fort Benning included in the final report (Table 17)

contains numerous errors that makes it difficult to analyze.

[43] The standard deviation predicted by the model should be close to the weighted L^1 or L^2 norm (depending on the multicollinearity of the variables) of the standard deviation divided by the square root of R^2.

[44] We can compose a composite mean score by summing the test event means multiplied by their respective standard deviations and dividing the quantity by the square root of the sum of the squares of the standard deviations. We can compose a composite standard deviation by taking the square root of the sum of the squares of the standard deviations.

[45] By taking the dot product of the coefficients and the means added to the constant term, we get a total of 248 seconds, which agrees closely with the average WTST hasty-fighting-position vignette mean of 268 seconds. Taking the dot product of the absolute values of the coefficients and the standard deviations, we get a total of 88 seconds, which agrees moderately well with the average WTST hasty-fighting-position vignette standard deviation of 67 seconds.

[46] The composite mean μ_* and standard deviation σ_* are computed using weighted z-scores $\mu_* = (\sigma_1\mu_1 + \sigma_2\mu_2 + \sigma_3\mu_3)/\sigma_*$ where $\sigma_* = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$ and μ_i and σ_i are the mean and standard deviations of the sled push, sled drag, and shuttle run.

[47] Malcolm B. Frost and Whitfield B. East, "AUSA 2018 Warriors Corner #5 - The Role of Army Combat Fitness Test," October 2018,

[48] Fort Benning data is not broken down by by gender in the BSPRRS final report, but overall it agrees with the Fort Riley data.

[49] Researchers tested participants using a five-repetition-maximum deadlift and converted then to the scores to an equivalent one-repetition-maximum deadlift using the Wathen formula: $w_1 = 100w_r/(48.8 + 53.8\exp(-0.075r))$ where w_r is the weight lifted for r reps. The ACFT uses a three-repetition-maximum deadlift. For the purposes of the table, ACFT standards have been adjusted to match the one-repetition-maximum scores listed in the BSPRRS final report. Because of changes from five-repetition-maximum to three-repetition-maximum using one-repetition-maximum, the scores listed in the table may not be representative.

[50] Information Paper, "FY19 ACFT Test Battalions—Additional RFI Answers," July 9,2020.

[51] Kyle Rempfer, "Soldiers can substitute a 2-minute plank after attempting the leg tuck on ACFT," *Army Times*, June 2020,

[52] *U.S. Army ACFT Field Testing Manual*, August 2018

[53] Personal communication with exercise physiology researchers Dr. Vince Tedjasaputra and Dr. Stewart Petersen.

[54] Many might argue that a leg tuck does npt replicate climbing up and over walls obstacles or exiting disabled vehicles either.

[55] The ACFT uses hand-release push-ups whereas the AFPT uses a traditional push-up. A hand-release push-up requires the chest to touch the ground and hands to be lifted up between repetitions. To some hand-release push-ups are more difficult and to others traditional push-ups are more difficult.

[56] The Army has not only reduced the minimum scores needed to pass, they also reduced the maximum scores. The ACFT maximum for push-ups is 60 repetitions, the same maximum expected of a 46-year-old male soldier under the old APFT standards.

[57] The Army expects run scores to be slower due to the other strenuous activity of the ACFT.

[58] Joseph Knapik et al., *Army Physical Fitness Test (APFT): normative data on 6022 soldiers*, U.S. Army Research Institute of Environmental Medicine, 1993

[59] Paul Devita et al., "The relationships between age and running biomechanics.," *Medicine and science in sports and exercise* 48, no. 1 (2016): 98–106

[60] Missy Ryan, "The Army is rolling out a new fitness test: Will it hold back women?," *The Washington Post*, September 2020,

[61] From Update from the SMA (Sergeant Major of the Army) July 2020 to U.S. Army Center for Initial Military Training: Trainees from the C/1-40 FA(BCT) took the ACFT on April 20, 2020 followed by the APFT on April 24, 2020. The number of soldiers who passed or failed are as follows:

	male pass	male fail	female pass	female fail
ACFT	98	3	50	45
APFT	44	68	31	46

[62] Gold-standard pass rates on the ACFT (from Army memo dated 09 July 2020 in response to Senator Gillibrand's request for information):

	FY19	Oct	Nov	Dec	Jan	Feb	Average
Female	21	37	29	33	34	40	32
Male	81	91	89	91	90	90	89

[63] Sean Robson et al., *Fit for Duty?: Evaluating the Physical Fitness Requirements of Battlefield Airmen* (The RAND Corporation, 2017)

[64] Chaitra M. Hardison, Susan D. Hosek, and Chloe E. Bird, *Establishing Gender-Neutral Physical Standards for Ground Combat Occupations: Volume 1. A Review of Best-Practice Methods* (The RAND Corporation, 2018)

[65] U.S. Army, "FM 7-22 Holistic Health and Fitness," *Washington, DC: Headquarters, Department of the Army*, 2020,

[66] Also see "How to be Less Wrong and More Useful: Thinking Critically about Decision Models", Mar 1, 2020

[67] Carl T Bergstrom and Jevin D West, *Calling bullshit: the art of skepticism in a data-driven world* (Random House, 2020)

[68] Public Law 115 - 435 - Foundations for Evidence-Based Policymaking Act of 2018

This report was written by Kyle A. Novak, Ph.D., a legislative fellow working in a personal office of the U.S. Senate. The fellowship was supported through a grant by the Alfred P. Sloan Foundation and sponsored by the American Statistical Association and five other mathematical-statistical societies to improve evidence-based policymaking in the federal government. The views expressed in this publication are solely those of the author.

Updated: November 19, 2020