



Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury (DCoE) Webinar Series

December 10, 2015, 1-2:30 p.m. (ET)

“Head to Head” Study: A Psychometric Comparison of Brief Computerized Neuropsychological Assessment Batteries

Good morning. Thank you all for joining us for the DCoE Traumatic Brain Injury December webinar, "Head to Head" Study: A Psychometric Comparison of Brief Computerized Neuropsychological Assessment Batteries. My name is Emma Gregory, and I'm a Research Scientist with the Defense Center and Brain Injury Center, and I'll be the moderator for today's webinar.

Before we begin, let's review some webinar details. If you experience technical difficulties, please visit www.dcoe.mil/webinars to access troubleshooting tips. Please feel free to identify yourselves to other attendees via the Chat box, but refrain from marketing your organization or product.

Today's presentation, references and resources are available for download from the Files pod and will be archived in the Online Education section of DVVIC's website. All who wish to obtain Continuing Education credit or Certificate of Attendance and who meet eligibility requirements must complete the online CE Evaluation. After the webinar, please go visit www.dcoe.cds.pesgce.com to complete the online CE evaluation and download or print your CE Certificate or Certificate of Attendance. This evaluation will be open through Thursday, December 24, 2015.

Throughout the webinar, you're welcome to submit technical or content-related questions via the Q&A pod located on the screen. All questions will be anonymous. Please do not submit technical or content-related questions via the Chat pod.

I will now move on to today's webinar, as I said, "Head to Head" Study: A Psychometric Comparison of Brief Computerized Neuropsychological Assessment Batteries.

With more than 300,000 service members diagnosed with traumatic brain injury, TBI, since 2000, a need for fast and easy assessment of cognitive functioning has arisen. Numerous computerized neurocognitive assessment tools, or NCATs, have emerged from this need. Companies creating these tests often tout them as suitable alternatives to traditional paper and pencil tests. However, emerging research suggests this may not be true, and the issue is not as straightforward as once believed.

Investigators at Fort Bragg, North Carolina, recently completed a two-phase study of the psychometric properties of four NCATs: ANAM4, CNS-Vital Signs, CogState and ImPACT. The first phase investigated the test-retest reliability of the NCATs by comparing examinees' scores over a 30-day interval. The second phase investigated the validity of the NCATs by comparing the performance of healthy service members and service members with acute mild TBI to their performance on traditional tests. In the context of this study, this webinar will present the state of the literature regarding NCATs, their clinical utility and future directions.

At the conclusion of this webinar, participants will be able to: identify key concepts when considering computerized neurocognitive testing as an alternative to traditional paper and pencil neuropsychological tests; describe the current state of the literature regarding computerized neurocognitive testing, to include recently-completed research at Fort Bragg, North Carolina; to articulate potential issues and future

directions with regard to evaluation of computerized neurocognitive tests for future use in clinical populations.

Now I would like to introduce today's speaker. First, Dr. Wesley Cole is the Senior Clinical Research Director and Neuropsychologist with DVBIC at Womack Army Medical Center at Fort Bragg, North Carolina. Dr. Cole completed his pre- and post-doctoral training in counseling and neuropsychology at the Kennedy Krieger Institute, which is an affiliate of Johns Hopkins School of Medicine. Dr. Cole holds a Ph.D. in Clinical Psychology from the University of South Carolina, an M.A. in Psychology from the University of South Carolina, and a B.S. in Psychology from the James Madison University.

Mr. Jacques Arrieux is the Senior Clinical Research Associate with DVBIC at Womack at Fort Bragg. His previous positions at Womack involved the administering and scoring of neuropsychological test as a psychometrist. And Mr. Arrieux has a Master's in Experimental Psychology from Fayetteville State University and a B.S. in Psychology from the University of North Carolina at Pembroke.

And now I'll turn it over to Dr. Cole to begin the presentation.

Thanks, Emma, I really appreciate it. Let me get my webcam here lined up.

First of all, disclosures – always important: This work was funded by DVBIC, in part, through contract support provided by the Henry M. Jackson Foundation for the Advancement of Military Medicine and General Dynamics Information Technology. The views expressed herein are those of myself and Jacques; and they don't reflect the official policy of the Department of the Army, Department of Defense or DVBIC. I'm not going to discuss any off-label or investigative use of commercial products or devices. And I have no relevant financial relationships to disclose. That business out of the way.

We wanted to go through a few polling questions first, just to get a sense of who the audience was. And this first question is just to get a sense of what disciplines are viewing this. So if you could, just take a minute to select which discipline best fits your category.

[Pause for responses]

It looks like we have a lot of rehab providers, behavioral health providers, and then people that fall into the "Other" range, great – yeah, about one-third behavioral health, about one-third Other, and about one-fifth rehab providers. That's great; I appreciate that.

Let's move on to the next question. This we were just curious as to who was listening in, in terms of what population you primarily work with. Do you work with active duty service members; do you work primarily with veterans; or if you work with civilians, are you working primarily in an athletic setting, so maybe like a sports medicine provider, or are you working primarily in a non-athletic setting; or do you work with civilians where you might see both athletes and non-athletes?

[Pause for responses]

The results are coming in; just over half are working with the active duty service members, and then about one-quarter with veterans, and then the rest civilians – great, thank you.

I think we've got one or two more of these. This is great just to give us a sense of who we're speaking with. Do you currently use or interpret scores from computerized test batteries? We're curious to see how widely these are used by our audience.

[Pause for responses]

We're getting about 40% to 45% are actively using these, and the rest are not – okay, great.

[Pause for responses]

It looks like we jumped back there. We've got too many people controlling the reins here.

All right, here we go, this last polling question: Are you familiar with any of the tests? These are the ones we're going to be primarily talking about. We're just curious if the audience is familiar with these. Just select if you're familiar with any of these; you can select all that apply. If you're familiar with another type of computerized cognitive test, go ahead and select that as well or if you're not familiar with any of them.

[Pause for responses]

A lot of people are familiar with the ANAM, which makes sense since we have a large active-duty audience. I have to say, I'm actually encouraged to see that there are about 30% not familiar with it; so that means you'll be exposed to some new information, so that's great – good, thank you.

I think we're at slide 16. Perfect.

I just wanted to begin after the polling questions by discussing where we're going to go. So this is our roadmap for the presentation. We're going to start by discussing a more general overview of neurocognitive testing; and specifically, computerized neurocognitive testing. As part of this discussion, we'll talk about some of the reasons using computerized neurocognitive tests are appealing to professionals.

And let me point out, for the sake of brevity, I'm going to refer to these computerized neurocognitive testing tools as NCATs. You'll also hear sometimes the term CNTs, or computerized neurocognitive tests; but the DoD refers to these tests as NCATs, so we'll go with that term.

Next we're going to discuss some of the potential drawbacks to using NCATs. And then we're going to provide you with a brief overview of our study to just give you some context for the rest of the presentation. This is going to set us up for an in-depth review of the four NCATs we studied, and this will include a crash course on the existing research of these NCATs. And then we're going to move back to our study and discuss the findings that we have to date. And finally, we're going to end with a discussion of the broader implications all of this has, to include the use of NCATs and future directions for research.

This is slide 17.

When we say neurocognitive testing, what exactly do we mean? I'm guessing most of the audience is familiar with neurocognitive or neuropsychological testing to some degree. Briefly what we're talking about are tests that are administered in a standardized way that are designed to measure various cognitive processes such as intelligence, attention, memory, executive functioning, among other areas of cognitive functioning.

Traditional neuropsychological tests are generally "pencil and paper," typically administered in a one-on-one, well-controlled setting with a neuropsychologist or trained psychometrist. The scores are then interpreted by a neuropsychologist or another trained professional qualified to interpret the scores. The tests used in neuropsychological testing have been normed; they've typically been well-validated and deemed reliable. We'll talk more about validity and reliability in a few minutes.

Now we're on slide 18.

What I wanted to do here is before we talk about computerized neurocognitive testing, I just wanted to put the larger issue into more context. Since Jacques and I are part of DVBIC, we tend to look at issues through the lens of brain injury. And really, it's brain injury, and specifically concussion or mild traumatic brain injury, that's driven the increased use of computerized tests.

So if you're a sports fan, and especially a fan of the NFL, you may have heard of athletes being tested for concussion. And in the NFL, they call that the concussion protocol. Part of this involves computerized

neurocognitive testing. In the military, there have been over 300,000 diagnosed TBIs since 2000, with the majority of those in the mild TBI classification. This does not include all the TBIs that were not reported or not diagnosed properly or otherwise not tracked.

So the increased awareness about concussion and the importance of adequate concussion care, the increased risk with these types of injuries with certain groups like athletes and military, and the high volume of concussions sustained by our service members, raised a need to be able to quickly and efficiently screen for cognitive functioning and specifically cognitive deficit after concussion.

Slide 19.

Out of this need that I just spoke about has risen the field of computerized neurocognitive testing. To be clear, using computerized tests is not necessarily new or specific to the need for screening for concussion. And, in fact, traditional neuropsychological testing can include computer-based tests; for example, the continuous performance test is an example of a computerized test.

NCATs are not used solely for concussion screening. However, the issue of concussion has really driven the increased development and use of NCATs. There are multiple reasons NCATs are appealing. First of all, they're often much shorter in duration from traditional tests. The four NCATs we studied range from about 10 minutes to 30 minutes in length to administer. And although the length of different versions and different tests can vary and they can be longer than 30 minutes, they're still generally much shorter than traditional tests, which can take hours or even be spread out over a few days.

You also don't need specialized training comparable to a neuropsychologist or a psychometrist to administer the tests. All that's needed is a computer testing session and a screening proctor. Large groups can be tested at once, such as in a computer lab. For baseline testing, whether it's a whole team or large groups of service members before deployment, this is a very efficient approach to testing.

There is some evidence that suggests group testing is not ideal, especially if post-injury follow-ups are going to be one-on-one or in smaller groups. But there is still some debate in the literature about that.

Another advantage is that the tests have the potential for almost unlimited alternate forms. Large banks of questions or stimuli can be randomly selected by the program to piece together similar and theoretically comparable alternative versions of the tests. So for repeated administrations, this is important to mitigate practice effects.

In traditional testing, timed tests rely on an examiner using a stopwatch. And therefore, there is some human error introduced. A computer can time things much more accurately and much more precisely, down to the hundredths or even thousandths of a second.

An NCAT can also rapidly produce results versus traditional testing, which uses hand scoring often where you have to look up scores in tables or enter raw scores into the scoring program to get your final standardized scores. With NCATs, the scores are often immediately accessible to the examiner. Additionally, many of the NCATs, as Jacques will soon show you, spit out organized reports with standardized scores.

And finally, many of the NCATs have centralized data storage, which can allow for analyses of large full datasets. It can also allow for norms to be frequently updated.

All right, moving on to slide 20.

So there is a fairly compelling case for using NCATs. However, they are not without drawbacks. Listed here are a few of the cons that have been discussed in the literature. The first is cost, which can include upfront costs for the required technology in the actual program as (inaudible) report costs that some of the NCATs charge. And if you're setting up a computer lab, the costs can go up very quickly.

Access can also be an issue. And what I mean by that is if testing is needed, you need access to a computer. And for some of the testing, you need access to a network. So if we think about the military, this may not always be possible in some settings. Or, if you want to make it possible, you'll need specialized equipment, like a ruggedized computer; and then the testing site may not be ideal.

Also, many features of the test are proprietary in nature; and therefore, scoring algorithms, norms and other features of the test may not be available to the examiner. We've run into this issue with some of our research; for example, getting the raw data that's used for calculating (inaudible) scores has proven difficult.

There are also hardware and software considerations. There's a great paper by Cernich and colleagues from 2007, and that's listed in our Resource Guide in our citations at the end of the slideshow. And this clearly lays out some of these considerations. So all of us use computers, and we all know not all computers perform the same; and not all computers perform consistently. Therefore, things like hardware configurations, background software, the type of mouse or keyboard used, and then how those are connected to the computer. And also, some NCATs use a keyboard for responses, some use the mouse for responses; but those can all affect the performance of the tests.

Additionally, some of the tests are network-based, meaning they're administered as a Web-based application. Therefore, and especially when measuring the timing of responses, all of these things can be an additional source of error and should be accounted for.

And then one-on-one testing with traditional tests, you're working with a trained testing expert. And you get a level of qualitative data that's not possible when testing is completed solely on a computer. These behavioral observations can be a key data point in a neuropsychologist's interpretation and conclusions that they make.

And you'll remember from the last slide that I mentioned precise measurements as a positive thing. Well, I believe it can also be a negative thing. Human reaction time is only so fast; I believe it's around a quarter of a second. If you want, go to www.HumanBenchmark.com to test your reaction to some fun exercises, www.HumanBenchmark.com.

There may be no real gain in measuring things down to the thousandths. In fact, we've hypothesized that within our data, these precise measurements may negatively affect some aspects of psychometric properties. And finally, those psychometric properties, specifically validity and test-retest reliability – they've really not proven to be as robust as we would like tests to be. And this has implications for clinical decision-making. And a larger amount of error than we would desire could be introduced into the results.

Slide 21.

How exactly are these NCATs used? Generally, they are *not* used to diagnose concussion and brain injuries. First of all, a brain injury is an event; and a follow-up assessment, such as cognitive testing, evaluates the symptoms resulting from that event. With civilians, ImPACT is the test that is most widely used, and it's part of many sports leagues' post-injury testing procedures.

Many teams in leagues conduct preseason baseline testing and then will conduct post-injury assessment, with the goal for the athlete to return to baseline levels before returning them to play. Outside of sports, these tests can also be used as traditional neuropsychological tests are used. Ideally, they would be used as a screening tool or a supplement to traditional tests rather than as a replacement for those traditional tests.

In the military, ANAM is the test that is primarily used, at least with the Army. However, some groups within the military use other tests, such as Special Forces uses ImPACT. They're used similarly to how they're used with athletes, except we're talking about pre-deployment or pre-baseline training rather than preseason training. And then post-injury evaluations have the goal of returning the service member to baseline levels before they return to duty.

Slide 22.

In 2011, the Defense Centers of Excellence and the Defense and Veterans Brain Injury Centers released clinical recommendations that were used in NCATs. That's provided as part of the Resource Guide that came along with this presentation. And on this slide is a screenshot of what the first page of that CR looks like, as well as some key points from the Guide. There are a few things that I want to highlight.

The CR states that these tests should be one component of post-injury testing and that they are not sufficient in isolation for return-to-duty decisions. They recommend testing parameters, including the ideal setting and the timing of testing. Specifically, the CR states that there are some individuals who aren't going to benefit from administration of these tests. They also recommend that the tests should be initially administered within 24 to 72 hours of injury and repeated every three to four days until symptoms resolve and there's a return to baseline.

They also recommend that a psychologist and a neuropsychologist are consulted for interpretation of scores.

Slide 23.

So that was your crash course on the general topic of NCATs. I briefly want to talk about our study, just to introduce you to it. We refer to this as the Head to Head Study because we investigated four NCATs and an homogenous example of service members. That is, we put these tests up against one another in a head-to-head manner – sort of; and we'll clarify what we mean by that.

The main goal was to evaluate the psychometric properties of the four NCATs and, specifically, test-retest reliability and validity. The four NCATs we used were the ANAM4, CNS vital signs, CogState and ImpACT. These tests were selected based on widespread use and the companies' willingness to participate and contribute their tests to this study. There were other tests solicited, but not included for various reasons.

And for transparency, I just want to tell you that those decisions were made well before I became involved in the study. And then once I took it over, I tweaked some of the procedures within the framework of using these four tests.

One of the changes I made was to break the study out into two phases or arms. The first phase focused on test-retest reliability, and then the second phase focused on validity.

Slide 24.

Before we discuss the results of the test-retest reliability phase, I just want to provide a quick reminder as to what reliability is. This refers to the test's ability to consistently measure a trait across time. So if a test is reliable and changes from one session though the next, they're going to be due to changes in the individual. Stated another way, if a stable individual takes a reliable test two or more times without a significant medical or other type of event occurring between the testing sessions, then the scores should be relatively the same.

A test with low or poor reliability has fluctuations in the scores that are due to testing error with various potential sources of that error. Changes in scores from one session to another cannot be confidently linked to changes in the individual.

One of the statistics most commonly used to evaluate test-retest reliability are correlations. A high reliability will be indicated by high correlations between the same score at two or more time points. The Pearson r correlation, which is a statistic most of us think about when we mention correlation is commonly used. Also, the intra-class correlation, or ICC, is also used. And the ICC has been proposed as a

potentially preferable score for test-retest reliability, as it's more ideal for within subject analyses; and that's the stat we used in our studies.

There are also other approaches, such as reliable change indices, effect size and competence interval based decisions, and regression-based approaches. But for time purposes, I'm not going to go into detail regarding those.

Slide 25.

The targets that you see are visual examples of reliability. Both of those targets indicate good reliability. As you can see with that top target, you could miss the center of the target but still be reliability. It goes for any test to have good reliability, so having all of your shots clustered together and also having your shots in the center, hitting the target. The descriptive terms for ranges of correlation coefficients are also on this slide, with tests involved in clinical decision-making preferably in the high or very high ranges.

Slide 26.

Let's also do a quick review of validity. Boiled down to its basics, validity refers to the test measuring what it claims to measure. So a valid memory test actually assesses memory and not some other trait. There are multiple types of validity, and we've seen them referred to with various terminology. So we're not going to get caught up in the semantics. We're just going to go to the terms that we have here on the slide and focus on the actual definitions because that's what drives what we're actually doing, how we're looking at the data.

Criterion validity describes a test's ability to assess a criterion that's external to the test itself – so medical diagnosis or classification or something like that. Construct validity refers to the functional relationships between the variables assessed by the test. I'll explain those in more detail in the next slide.

Let's look at the targets first. The top target has the shots scattered all around, which is representative of a test with poor validity. It also has poor reliability because the shots are clustered together. The bottom target has the shots clustered in the middle, and that's indicative of good validity; that is, our test is hitting the target by measuring what it says it measures. You'll notice it's also reliable since the shots are clustered.

So here's where I'll note an important point. Reliability is a necessary but not sufficient condition of validity. So remember, you can have a reliable test that's not valid; but you can't really have a valid test that's not reliable. This poses some pretty big challenges in our interpretations, which we'll get into later.

We're primarily interested in criterion and construct validity, which I previously mentioned. We'll look at concurrent validity, which is comparing the test to an existing benchmark or gold standard, such as pencil and paper neuropsychological tests.

Predictability will refer to how well a test predicts future abilities or medical diagnosis or classification. Convergent validity looks at how well related abilities hang together. So is a memory test hanging with other memory tests or with test and processing screens? Tests with discriminant or divergent validity would be unrelated, which would be expected if they are tests of different cognitive abilities. So a memory test and a test of processing speech have little to no relationship, indicated by being relatively uncorrelated.

Slide 28.

After hearing validity described, you can probably imagine that there are numerous methods for evaluating it. We can look at correlations, and we can use those correlations to look at the relationship between NCATs and traditional tests or other tests with similar abilities. We can use regression analyses to see how well NCATs predicts certain traits; that is, what amount of variance do NCAT scores account for in a predictive model?

We can also use statistics referred to as "classification analyses." This refers to things like Receiver Operating Character Curves or ROC. That's essentially a plot at how accurate your measure is at identifying some classification using a range of different criteria, ultimately and ideally allowing you to select the optimal diagnostic cutoff.

There are sensitivity and specificity analyses, which tell you how well your measure is identifying people without that certain condition. There are positive and negative predictive values, which take into account the sensitivity and specificity, as well as the prevalence of a condition and tell you the ratio of true positives and true negatives to overall test results. Odds ratios tell you how likely someone with a certain characteristic or test score would be categorized or diagnosed with a certain condition and so on.

So in other words and with regard to NCATs, we can see how accurate NCATs are at classifying patients into specific categories, such as history of TBI, presence of cognitive deficits in veterans.

And finally, we can use approaches like principal components analysis or canonical correlation analysis. This looks at how NCATs scores load on various components. Those components can be defined by predetermined categories, traditional test scores or other statistical methods. These approaches can give us an idea of what traits NCATs are actually assessing.

Moving on to the next slide.

Credit to Jacques for finding this definition: "All of these labels for distinct categories of validity are ways of providing different types of evidence for validity and are not, in and of themselves, different types of validity as some sources might claim." In other words, the labels or categories, and even the analyses, are different ways of looking at the same thing, which is validity.

Okay, so here I'm going to give you a break from listening to me; and I'm going to turn things over to Jacques. There was an introduction done for Jacques. He is our Senior Clinical Research Associate here at DVVIC. He's been intricately involved in this site, and I'm going to let him take over from here.

[Pause for change of speakers]

Hi, everybody. My name is Jacques Arrieux, and I'll be presenting on the different NCATs and some of the psychometric properties of these NCATs.

I have no relevant disclosures to inform you of.

For ANAM, this is the most commonly used NCATs in the DoD. It's typically administered to service members before and after deployments; and the ANAM is similar to traditional neuropsychological tests. It provides measurements of cognitive processes that are typically affected by mild TBI.

The ANAM4 military battery consists of seven primary subtests. And these subtests are listed on the left slide of this slide, and there are some screenshots of the subsets listed on the right. These subtests measure cognitive function, such as processing speed, attention and memory; and performances on these subtests is measured by the clicking of a mouse, which generates scores for reaction time and accuracy.

For example, I'll describe the Code Substitution – Learning subtest, which is the second screenshot down. Some of you will notice that this task looks very similar to the way single-digit coding task. In this subtest, the examinee is to click left on the mouse if the digit pairing correctly matches the key above; and if the pairing is incorrect, then the examinee is to click right on the mouse. So for time's sake, I can't go into great detail about all of the subtests, but this should give you a good idea of the look and feel of the ANAM.

Overall, it takes about 30 minutes to complete the seven subtests in the ANAM4 Military Battery. However, the ANAM does have a library of several other subtests that can be used to create various batteries; but, again, the ANAM4 Military Battery consists of the core seven subtests.

Across the seven subtests on slide 34, you'll find across these seven subtests the performance is typically evaluated with a throughput score, which basically takes into account both speed and accuracy into the consideration for the interpretation of performance. The overall score is defined by the average performance across the seven subtests and is then computed into a composite score. And also in pair performances calculated as a score, that's less than 1.28 standard deviation from the mean.

On slide 35, you'll find an ANAM Performance Report which breaks down ANAM scores – and this is a de-identified report, so the nine-digit number there is a randomly-assigned number. And on the right side of the report, you'll find reaction time, percent correct and throughput scores for each of the subtests. You'll also find the raw scores, the percentile scores and the standardized scores for each of the subtests; and also, there's a comparison group listed on the far right column.

The ANAM, as with the other NCATs, also provides data exporting features so that the data listed on this report can be dumped into an Excel spreadsheet for further statistical analysis. Also, the ANAM as well as the other NCATs provide a validity indicator to provide some information regarding the examinees' level of effort for the credibility of the assessment.

On slide 36, I'll move into a review of the psychometric properties of ANAM. First, I'll review the test-retest reliability studies. In this table, you'll find the author, the sample, the test-retest interval, the statistical analysis and the results listed under the test-retest section. And I want to draw your attention to the results.

Although in the seven studies that we reviewed, the reliabilities were variable from low to high, most of these studies have found moderate reliability at various test-retest intervals with healthy controls. You'll also find the Cole et al study that we conducted here at Bragg highlighted, and we'll talk more about the results of that study later on in the presentation.

Over the next few slides, I'll summarize some studies that evaluated ANAM's evidence for validity. My intention is not to go into great detail for every study because a lot of the pertinent information on these slides here is for reference for those of you that actually have a physical copy of the slideshow. You're encouraged to return back to the actual manuscript for additional information about any particular study. Since there's so much data and so many new studies popping up every day, my goal is to highlight some of the key studies so that we can understand the various accessible approaches to investigating the psychometric properties and to gain a cursory knowledge of the existing literature on each NCAT.

On slide 37, I've summarized three studies: by Woodhouse and colleagues, by Kelly and colleagues, and by Register-Mihalik and colleagues. And there are a couple of key things to focus on in this and subsequently similar tables.

First, the methods – as you can see, the three studies on this table looked at the validity of ANAM in slightly different ways, either by comparing mTBI to controls or to neurologically impaired folks to controls. The statistics used were logistic regression, receiver operating characteristics or ROC curves, reliable change indices or RCIs. And these studies generally evaluated the accuracy of classifying individuals into the appropriate groups.

So what I really would like you to focus on are the results. And the stats aren't generally consistent across these three studies. But the trend is that specificity, which ranges from about 82% to 98%, is stronger than sensitivity, which ranges from about 9.3% to 81%, largely dependent upon the inclusion of different sources of information, like postural stability and symptom report.

On slide 38, we'll summarize three more studies that looked at the validity of ANAM, again drawing your attention to the varied procedures and the highlighted results. First, in a study by Norris and colleagues,

they found that simple reaction times two, or simple reaction time repeated, was an adequate predictor of recovery time.

In a second study by Coldren and colleagues, they found that the ANAM – you can adequately distinguish between those with mild TBI and control during the acute period; but after about five more days, there was little of important difference.

And finally, in a study by Bleiberg and colleagues, they used principal component analysis, or PCA, to identify four factors of ANAM with the traditional battery. They also found with stepwise regression that the mathematical processing subtest and the Sturmburg search test were good predictors of traditional scores.

In slide 39, you'll find four more studies that looked at the validity of ANAM using various methods. And the first two studies listed there, by Jones et al and Kabat and colleagues, they found three factor solutions that accounted for over 60% of the variance of ANAM and traditional test.

Next, Guskiewicz and colleagues compared the ANAM's rating of impairment to other impairment measures and found 22% to 52% disagreement rate.

And finally, Hawkins and colleagues found that the ANAM correctly classified over 83% of individuals as impaired and 86% of individuals as not impaired.

On slide 40, you'll find a study by Meier and colleagues from 2015 that compared 17 college football players with acute concussion to health controls. They found significant differences on Simple Reaction Time tests 1 and 2, which was also associated with cerebral blood flow imagining and a longer recovery time. So we thought that was an interesting study to include in the review here, and that just recently came out in *JAMA*.

On slide 41, you'll find five studies that compared traditional neuropsychological tests to the ANAM using correlations. And, as was mentioned, this is one of the most common ways of evaluating the evidence of concurrent or convergent validity between well-established tests and tests whose psychometric properties are currently being evaluated.

And the results were generally variable and not what one would expect from purportedly similar measures. So overall, the correlations were moderate at best. And again, my goal is to expose you to the general body of literature on ANAM, describe some of the ways that reliability and validity have been investigated. And this will put our results and our methods into more of a context. And I'll provide similar overviews for the other three NCATs.

So next up is CNS vital signs. The CNS Vital Signs is a more commonly used test in amateur athletics, clinical trials, and different psychiatric settings. It measures different cognitive functions that are commonly affected by TBI, other neurological and psychiatric issues.

On slide 43, you'll find that the CNS Vital Signs, a battery that we used, consists of seven subtests, of which three of the subtest screenshots are listed on the right side of the slide. The test battery includes immediate and delayed memory subtests for verbal and visual information. There are measures of executive functioning, processing speed, attention and so on. The examinees are to use both mouse and keyboard to respond to stimuli on various subtests, and performance is measured by standardized subtest scores, domain scores and an overall index score.

On slide 44, you'll find the 11 domain scores and the overall composite score that's listed on the left, and the formulas for calculating the domain score is listed on the right side of the slide.

Move on to slide 45, and you'll find another de-identified report. And this is one of the clinical reports for CNS Vital Signs. And this is kind of half of the report. This part of the report provides a nice graphic of the

performance across the 12 domains. It provides raw scores, standardized scores, percentiles and a validity indicator for each section.

On slide 46, you'll find the second half of the report, which gives you some subtest performance. And in this part of the report, it also gives you raw, standardized percentile scores for each subtest. In addition to that, it also gives you a description of the subtest, which is on the far right column of the report.

On slide 47, I'll move into a review of the psychometric properties of CNS Vital Sign. First is test-retest reliability study. The Cole et al study conducted here at Bragg is highlighted here in yellow, and we'll discuss more of those results later on in the presentation. But as you can see, the results are similar to the other NCATs in that the reliabilities are variable, from low to high. And it could be based on probably the test-retest interval and the group of individuals that were being evaluated.

On slide 48, you'll find a table summarizing four additional analyses, which evaluated the validity of CNS vital signs. First, in a study by Gualtieri and colleagues, they found significant differences between healthy controls in patients with psychiatric disorders. Next, Lanting and colleagues found non-significant differences and small effect sizes between participants with mild TBI in orthopedic controls. However, the mTBI group were observed to have more scores less than one standard deviation below the mean.

And in the last two rows of slide 48, you'll find a study by Gualtieri and colleagues. And what they found was when they compared scores based off of demographic factors, such as age and education, it contributed to differences. Also when they ran comparisons between healthy controls to an ADHD group and a TBI group, they found that neither stepwise regression nor logistic regression were able to identify a specific pattern of responding.

On slide 49, you'll find three more studies that evaluated the correlations between CNS Vital signs and a traditional neuropsychological test battery. All of these studies revealed moderate correlations at best. The comparisons are not always so straightforward. But notably, the 2006 Gualtieri study found the highest correlation was between the WAIS symbol-digit coding subtest and the CNS Vital Signs symbol-digit coding subtest. And this is what you would expect from correlations between traditional neuropsychological tests and NCATs, especially similar tests.

So correlations can be sometimes difficult, especially with dissimilar subtests. And in the Lansing study, this is a good example – they found that the highest correlation in their traditional battery compared to CNS Vital Signs was the NAB Memory Index and the CNS Vital Sign psychomotor speed and not CNS Vital Sign Memory Index. So I think this is a good example of how these comparisons just generally don't provide some results that you would typically expect when comparing similar tests.

Next up we're going to move into CogState on slide 50. And generally, CogState is a test used in Australia in athletic and military settings; and it's quite different from the other NCATs because it consists of only four subtests, which use a playing card motif. The four subtests could be administered in about 10 minutes. And on slide 51, you'll see a list of the subtests on the left; and you'll see some screenshots of the stimuli on the right.

The Detection Task is a simple reaction-type subtest which examines the examinee's ability to respond when the card turns over. The Identification Task measures the examinee's performance to determine whether the color of the card is red or not, and it kind of taps into processing speed. The One Card Learning Task, or OCL Task, is a memory task which presents cards to examinees; and they're asked to determine whether the card has been presented before. And lastly, the One Back Task subtest presents cards to examinees and records their responses in determining if the card is identical to the one shown before. And it seems to be tapping into working memory.

Moving on to slide 52, you'll find the primary scores, the outcome measures and the calculations for suboptimal effort, also known as data integrity. The performance is measured by both speed and accuracy, and the overall composite score consists of the average across the four subtests. The composite score uses speed for the calculation of the overall composite score, with the exception of the

OCL subtest, which uses accuracy. So there's a combination of both speed and accuracy scores that contribute to the overall composite score.

In addition, CogState provides a calculation for possible impaired performance. And CogState also has a large normative reference group, which includes age- and gender-based norms, which helps us to generalize standardized scores with the z-score. And we get all this data dumped out into a data export tool in a nice, neat Excel file for data analysis.

Moving on, on slide 53, you'll find that I summarized the test-retest reliability of seven studies. You'll notice that the results in the column of the reliabilities are variable, from low to high, depending on the test-retest interval. And some of these studies found that the reaction time scores were more reliable than the actual accuracy scores.

You'll also find the Cole et al study highlighted in yellow, and we'll discuss more results about that study later on in the presentation.

On slide 54, you'll find that I've summarized three studies that evaluated the evidence of the validity of CogState in different ways. The first two rows are studies by Maruff and colleagues in 2009 and Collie and colleagues from 2003. They found low to moderate correlations between healthy controls as performance on CogState compared to traditional paper and pencil neuropsychological assessment batteries.

The Maruff and colleagues study found that the highest correlation was between CogState's OCL, or One Card Learning Task, and the Brief Visual Special Memory Test, with the correlation of 0.83, which is great.

Jumping ahead to Louey and colleagues in a study from 2014, they demonstrated that controls performed better than acute mTBI on CogState. And in this study, they basically demonstrated to have good sensitivity, specificity and correct classification rates in identifying abnormal performance when using baseline scores and normative data. But the use of baseline data did increase sensitivity, specificity and correct classification rates.

On slide 55, you'll find that I've summarized four other studies that evaluate CogState validity with different methods. In these studies, they demonstrated that mTBI and other neurologically-impaired groups did not perform as well as controls on CogState. Notably, the Maruff and colleague study from 2009 demonstrated that performance on CogState differentiated between controls and clinical groups with these non-overlap statistics that they used. And they found that the OCL, or the One Card Learning Task subtest, accounted for about 78% of the difference between the mTBI group and control groups, which is good. We like to see those types of differences.

So on slide 56, I'm going to jump into ImPACT and give you a quick description of it, and we'll review the psychometric properties. ImPACT is probably the most widely-used NCAT out there. It's used in amateur and professional athletic settings, and it's also used in some of the DoD Special Operation Units. It consists of about eight subtests and takes about 30 minutes to complete. You'll find a list of the subtests on slide 57; and on the right side, you'll find a few screenshots of some of the subtests.

There's the Word Memory subtest, there's the Design Memory subtest and the Symbol Match subtest; they're shown on the right side of the screen. And again, ImPACT, as well as the other NCATs, generates standardized scores. In this case, they produce five composite scores, which are typically composite scores that you would expect to see in a typical neuropsychological evaluation.

And on slide 59, ImPACT also does generate a clinical report. And again, this is a de-identified clinical report that provides an overview of performance on the subtests and on the composite scores. The composite scores are compared to a normative reference group and then generate standardized percentile scores as well. The ImPACT test also provides information regarding participants' efforts, and it

also provides an option to export the results into an Excel file, which makes things a lot easier for data analysis.

On slide 60, you'll find an overview of the test-retest reliability of 10 studies. And like the other NCATs, the reliability coefficients reported were variable, from low to high. Most of these studies used different samples, different test-retest intervals, and different number of administrations, which can probably account for some of the variability between those reliability coefficients. You'll also find the Cole et al study highlighted in yellow, and we'll talk more about that in just a moment.

On slide 61, it provides an overview of selected studies that evaluated the correlations between ImPACT and traditional neuropsychological tests. Overall, the results were variable, with correlations ranging from low to adequate. Of note, Maerlender and colleagues in 2010 found that when traditional neuropsychological test batteries were compared to ImPACT with canonical correlations, the statistics demonstrated stronger correlations between the batteries.

On slide 62, you'll find that I've summarized three additional studies that evaluated the validity of impact. Of note, Maerlender and colleagues in 2013 ran additional analyses on that same dataset from 2010; and they made comparisons between the ImPACT battery and the traditional neuropsychological assessment battery. And these findings were interesting because although in the 2010 study the batteries were shown to demonstrate adequate evidence of convergent or concurrent validity, the 2013 study demonstrated that ImPACT's domain scores were correlated with dissimilar measures, which is not what you would expect from a traditional battery. So, in fact, you'll see with this study, it found that the traditional neuropsychological battery domains were not correlated with dissimilar measures, as you would expect to see with well-validated measures.

On slide 63, you'll find two more studies that evaluated the sensitivity and specificity of the ImPACT battery. Schatz and colleagues in 2012 found 91% sensitivity and 69% specificity of group prediction and membership.

Broglio and colleagues in 2007 evaluated ImPACT sensitivity to impairment, and they found ImPACT to be about 80% sensitive. And when cognitive data was combined with other information, like postural control and symptom assessment, the sensitivity increased to about 92%.

So you're encouraged to return to the actual manuscripts that I've covered here for additional information about any of these studies. There is so much data out there, and there are so many new studies popping up almost every day that my goal was to highlight some of the key studies so that we can understand the various accessible approaches to investigating the psychometrics of these different NCATs and to gain a cursory knowledge of the existing literature.

Now I'll pass it back over to Dr. Cole to discuss some of the results of the study that we conducted here at Fort Bragg.

Thank you, Jacques.

Let me get my Webcam. Note to self: next I hit pause, can the Webcam be more flattering. There we go, back up.

Moving on to slide 64, earlier I introduced you to the basics of this study; now I'm going to go into more details about the study design and the results. I'll begin by reviewing the test-retest reliability study. I'm going to point out that the results from this study had been previously published – I've mentioned that before – and that's listed in our references and on the Resources Guide. Then we're going to go through the design and the results from the validity study. And finally, we'll end with some general conclusions and broader statements about what this all means.

Going on to slide 65, for the test-retest reliability phase, we only investigated healthy controls. They were randomly assigned to take one of the four NCATs, and then we had them return approximately 30 days

later to retake that same test. Although 419 service members were enrolled, only a little more than 50% returned for the retest session. And this resulted in about 50 to 55 participants taking each NCAT at both time points.

Each NCAT has an effort indicator built in to ensure the test taker is putting forth an adequate effort and therefore giving good data during testing. After those individuals deemed putting forth poor effort were removed from analyses, we had 50, 39, 53 and 44 individuals for ANAM4, CNS, CogState and ImPACT, respectively included in the final analyses. It's important to note that we still had adequate statistical power for analyses. Also, the groups of soldiers taking each NCAT were statistically equivalent to each other across multiple demographic (inaudible).

Slide 66, initial analyses were conducted by a third-party statistics company that was blinded to the NCATs and the variables, just to ensure complete objectivity. Again, we excluded participants deemed to have put forth poor effort. And here's a description of how we calculated the ICCs. This was essentially a ratio of between person variability to the total variance. And again, ICC is good because it really takes into account (inaudible).

We also conducted reliable change indices based on a 95% confidence interval to determine if a higher rate of scores than would be expected, in this case approximately 5%, showed significant change from Time 11 to Time 2.

Slide 67.

These are the ICC results for ANAM. The ICCs are color coded with a key at the bottom. So anything in green is very high. In the high range is blue; adequate range is black; the marginal range is that yellowish-orange color; and then the low ranges are red. As you can see, the highest correlations were with the math processing and code substitution learning at 0.70 and 0.79, respectively.

Slide 68.

These are the ICCs for CNS. There are four scores in the adequate range: psychomotor speed, reaction time, complex attention and then the overall neurocognitive index. And forgive the formatting error there. When those top three scores are blown up, the second one there is speed and then a (inaudible) of two lines (inaudible); but there are four of those scores that are in the adequate range.

Moving on to slide 69, we've got the results for CogState. Four of the five scores are in the adequate range: detection speed, identification speed, one back speed, and then the total composite score.

Then moving on to ImPACT, slide 70, only one score was in the adequate or higher range; and that was visual motor speed, which is actually in the high range at 0.83.

Let's just summarize all of this. This is a summary table in slide 71. Many of the scores were in the low to marginal range, although each NCAT had at least one in the adequate or higher range, with CogState having the highest proportion in those ranges and ImPACT the only test with a score in the high reliability range.

Slide 72.

To summarize, generally we found test-retest reliabilities that were consistent with the other studies of test-retest reliability in the same NCAT. So Jacques reviewed much of this literature, and then I've provided some references that have overviews of the test-retest literature. And a more comprehensive list is in our references at the end of the slideshow.

When other studies have found differences, it's often due to significantly different test-retest intervals. For example, a study with only a few days to a week between test sessions may find better test-retest

reliability than studies, such as ours, with longer intervals or studies that have intervals up to a year longer.

What we did find is that test-retest reliabilities were generally not in the high or very high range. Again, I've already summarized how all NCATs had at least one subtest index or composite score in the adequate or higher range; and CogState had the highest proportion in the adequate or higher range.

What we found was that measures of response that actually tended to have the highest reliabilities and are consistent across and that the reliabilities were lower than desired for clinical decision-making. And again, we've published those results and archived the clinical neuropsych in 2013.

I also just want to point out, it's important to note that these results were not sufficient to identify a best test among the four NCATs. If you'll remember the targets, we did not want to run the risk of selecting a test that had all of its shots clustered but up on the edge of the target. So to identify best tests, we needed to investigate validity as well.

Moving on to slide 73, let's talk about that validity phase. We had two comparison groups: healthy controls in individuals in the acute and mild TBI phase, meaning they were tested within seven days from their injury. And the median time from injury was about four days. I do want to point out, we really would have liked to have done all of the testing within 72 hours or three days. For logistical reasons, it was just really difficult to make that happen. We strive to make that happen whenever possible, and so we got that median down as close to three as possible. That's just something I wanted to point out – that we were aware that getting them within three days would have been ideal.

For most of the analyses, we're looking only at enlisted personnel only; and this was to keep the groups (inaudible). We had a lot higher rate of officers in the control group, and that's likely due to the greater control officers have over their schedules. They're more able to come in and participate in a four-hour long study, more so than lower enlisted personnel might be.

But even with enlisted only, we had 139 controls and 216 in the TBI group. And this ended up resulting in over 50 participants taking each NCAT; and often we had more than that because we assigned participants to take two of the four NCATs. It was random assignment; and then we also counter-balanced the order of administration those NCATs were given. What I mean by that is sometimes the ANAM was first, the (inaudible) was second, or the CNS was first, second and so on.

We also administered a traditional neuropsychological test battery, always in the same order and always after the NCATs. The specific measures are listed there on the slide, but what we tried to do was develop a battery that looked at the major cognitive areas using measures that are well-validated and commonly used in clinical practice and then assess the same kinds of areas that the NCATs claim to discuss.

I do want to note here that we excluded participants who did not put forth adequate effort, as indicated on effort measures in the traditional battery or effort measures at (inaudible).

Moving on to slide 74.

Currently, we've broken the analyses out into four major approaches. The first was to investigate the affect that order of administration had on NCATs. This rules out a question I had about the necessity of counter-balancing the order. The literature that exists on traditional neuropsychological tests generally suggests that the order of test administration within a battery is largely a non-factor, and that's especially relevant with NCATs because NCATs are pretty self-contained.

What I mean by that is there's not a piece that you administer and then a delay where you might have missed another test and then you revisit that first test to do the delayed recall portion or something like that. You give the ANAM; it's self-contained. And then you give the CNS Vital Signs, and it's self-contained. There's no overlap there. So the literature on testing that exists around traditional tests would generally suggest that order of administration doesn't really matter. But we wondered if it did. So what we

did was we compared NCAT scores at time 1 to the same NCAT scores at time 2, and then further broke it down based on which test was received at time 1. I'm going to go into more detail on that later.

We also correlated NCAP scores with the traditional test scores, so we wanted to see if the scores from the NCAP kind of hung with similar scores from the traditional tests. Again, the NCAP memory scores correlate more highly with traditional memory test scores. We also used T-tests to compare controls in mild TBI participants' performance on the test.

Finally, we used logistic regression to look at the test's ability to predict cognitive impairment. This allowed us to control for various demographic factors. And cognitive impairment was defined by performance on the traditional test.

Moving on to slide 75.

Remember, these are between subject analyses for order effects, since each individual participant received two different NCATs. So they had only received a single NCAT at one of the two time points, first or second. For order effects analyses, we found no differences between ANAM scores at time 1 or time 2 or impact scores at time 1 or time 2.

We found potential slight to moderate order effects for CNS Vital Signs in CogStates, with some CNS scores slightly better at time 2 and some CogState scores actually slightly worse at time 2. However, once we controlled for false discovery rate, which is necessary given the large number of comparisons we did since we're comparing each NCAT scores, which is anywhere from 4 to 12 scores, several of the comparisons, especially for CNS, were no longer statistically significant. However, the effect sizes did remain somewhat moderate.

And what we kind of felt like and the other point I want to make is when we control for which tests they received at time 1, the results really didn't change significantly. So though there was not a universal order effect, there may be a slight order effect for some scores. So what we feel this justifies is accounting for the order of administration in these types of studies. They can either do that statistically or through study design, like counter-balancing the order of administration.

Moving on to slide 76.

This is a table that presents the correlations between the NCAT scores and similar traditional test scores. For example, we took the ANAM score that is supposedly a measure of processing speed and we compared it to the traditional test score that is a measure of processing speed. When there was more than one score to compare, we've reported the range of correlation coefficients. So we didn't want to make your eyes cross by presenting the huge correlation measures that this spit out.

These are all based on a-priori comparisons and, again, guided by the cognitive area each test claims to measure. So what I'm going to do is highlight – I've got bolded and circled correlations that are in the medium or approaching the medium range. And you can see, there's really no overwhelming patterns suggesting these are highly related. The correlations are all over the place.

Where you see a negative correlation by the way, that's just a function of the scores being on an inverse scale from one another.

When we looked at the correlation measures between supposedly unrelated cognitive domains – for example, the CNS attention test and compare that to the traditional neuropsychological visual scanning test and so forth – we saw a similar pattern, with a scattering of median correlations and no pattern of them being consistently unrelated. Again, we would have expected related tests to be highly correlated, unrelated tests to be lowly correlated. And we didn't really see that consistent pattern.

Now let's look at the comparisons on slide 77 between controls in participants with acute and mild TBI. I want to orient you to this table because it's going to repeat four times. This one is for ANAM scores.

Scores listed are the throughput scores for the ANAM for subtests. We have the mean and standard deviation for the controls; then beside that, we have the standard deviation from the mild TBI group; then we have the T-statistics for that comparison, then the approximately-value and then the Cohen's *d* effect size. To make it easier, we've color coded statistically significant findings in yellow, small effect sizes is green, medium is orange and large is red. The key is at the bottom. This format is going to repeat for each NCAT.

So for the ANAM, you can see all but one test was significantly different. And then post-(inaudible) analyses showed that controls performed better than those with mild TBI. But we saw small to medium effect sizes here.

Moving on to slide 78, where we have the comparisons for the CNS Vital Signs. Again, you see all of the composite scores that we compared listed there on the left, and then the scores in the table. And actually, you can see all tests were significantly different; and again, the controls performed better than mild TBI and we generally had medium effect size for CNS.

Moving on to slide 79, the top table is for Cog-State; and there you'll see four of the five tests were significantly different. Again, controls performed better than mild TBI. And for this, we primarily had large effect sizes.

The bottom table is for ImPACTs, where only one score was significantly different with small effect sizes for the comparisons.

Here is a summary of what we just went through. So 7 of the 8 ANAM subtests were different; all 12 of the CNS index scores were different; 4 of the 5 CogState subtests were different; and 1 of the 4 ImPACT indices were different, with CogState demonstrating the largest effect sizes. And in all cases, the controls performed better than the mild TBI group.

Now what we're going to do is look at the ability to predict cognitive impairment, with circles. I apologize; this was a popup, so it kind of mucked up the table there but I'll point out what we're talking about here.

Cognitive impairment – this is a bit arbitrary in that we had to come up with a way, based on the traditional neuropsychological test battery, to define impairment. We kept getting asked, did the NCATs predict impairment? When I hear that question, I don't really know how to answer it because how do you really define impairment? It's really based on are you talking about performance on a particular test – a memory test, an intelligence test, attention test? Are you talking about functional impairment, impairment defined statistically? Are you talking people in the borderline impaired range, in the impaired range? There are a million different ways. You could put 10 neuropsychologists in a room and get 11 different answers for how to define cognitive impairment.

But we had to go with a way, so what we did was we defined it as scoring less than 2 standard deviations below the mean on any one of 26 scores of interest from that traditional battery. So we pulled 26 scores that we felt were kind of the most robust or commonly interpreted scores. And if you were less than 2 standard deviations on any of those, then we called that cognitive impaired. We thought that was fairly conservative criteria, but it really wasn't.

We had about 75% of the total sample classified as impaired – 65% of those in the control group and 86% of those in the mTBI group. I think moving forward, we should reconsider our definition of impairment.

For the NCATs, we used their definition of cognitive impairment, either as defined in the manual or the general parameters or in criteria developed in conjunction with them. As you can see, the prevalence rate was between 37% to just under 45%, so largely in the same ballpark with each other. The sensitivity – referring to correctly classifying someone who was impaired – varied from CNS at the lowest, at 43.6%, to ANAM at the highest at 66.7%.

Specificity, the ability to correctly classify not impaired, ranged from CNS at 84% to CogState at 88.7%. When we look at positive predictive and negative predictive value, which take into account the sensitivity and specificity of the measures, as well as the base rate of the condition to determine how accurately we could predict someone as impaired or not impaired. We can see that the NCATs range from CNS in the mid to upper 60% accuracy rates and the ImpACTs in the upper 70% to 80% rates.

Slide 82.

Let's just summarize all of the validity results. We saw small to medium correlations with traditional tests, even among similar cognitive domains. With no clear pattern, similar domains were more highly correlated than dissimilar domains.

With the exception of ImpACTs, healthy soldiers consistently performed better on NCATs than those with acute and mild TBI. And this is in a manner consistent with traditional tests. However, the clinical meaningfulness of those differences is varied, with CogState the only one that had consistently larger effect size.

And ImpACTs seemed to predict cognitive impairment, as we defined it, relatively well with ANAM4 performed the best.

Let's pause for a moment; let's catch our breath. Remember the roadmap that I displayed at the beginning? We're almost at the last stop. So you're on that road; you can see your exit; you have just a few more slides left. So congrats and thanks for those of you who are still with us.

What I wanted to do was talk about some of the limitations and critiques of our study. First, there was a relatively small n for the reliability phase. Though we enrolled over 400 participants, only a little more than 50% returned; and that group was then divided among four NCATs, and additional folks were dropped due to inadequate effort. However, we were sufficiently powered for analyses, and our numbers were still in line with other reliability studies. And that's pretty evident from the numbers that Jacques put in his summary slides.

We also used the same computer platform for all four NCATs. Not all NCATs are designed to be administered the same way on the same platform. For example, CogState and ImpACTs are Web-based tests; and ANAM4 has specific platform guidelines that they provide. However, for logistical reasons, we had to run them on the same platform; and we received input from each test company on how to set up our computer platform.

We also used any of their recommended post-hoc data correction tools or approaches to adjust for potential hardware or software issues. And finally, we call this a "head to head" study, but we really didn't do a single analysis that compared all four n in a head to head manner in a single subject. In fact, we actually didn't present any analysis that looked at performance across two tests within a single subject. So we're not trying to fool you; we recognize and acknowledge that.

There are a few points to make. First, giving four tests to an individual would be difficult from a participant burden standpoint. Broglio and colleagues did a head to head study with three NCATs, and they were actually criticized by some other authors for administering more than one NCAT due to the potential effects it could have on performance in subsequent scores.

Also, the groups were statistically equivalent; and therefore, we feel that what we're doing is comparing these in a head to head manner in a homogenous sample. But we do have plans to take a deeper dive into the data and do some true head to head comparisons in the future.

Slide 84.

Some broader conclusions from our study – a "so what" wrap-up, if you will. Test-retest reliability is lower than desired, but our findings are generally consistent with the rest of literature. As you remember, liability

is a necessary condition of validity. So this hampers further investigation of validity to a degree. In those investigations, we saw a relatively poor convergent of validity. However, there does seem to be some potential utility of these tests, despite those results, at distinguishing between controls and injured patients, with that degree of utility varying by NCAT.

Also, there may be some clinical utility at identified patients as impaired or not, although the stats suggest the tests are better at identifying those patients that are not impaired. Again, the utility seems to vary by NCAT.

Now to our so what questions. Currently, we can't pinpoint a "best test" despite being asked this question; and I'm actually not sure that's the right question to ask. It's like asking which pencil and paper memory test is the best. As long as it's psychometrically sound, there are strengths and weaknesses of each one.

The questions should be centered around the purpose of using the test. Which test best fits similar goals? Are those goals to distinguish between someone who was recently injured, screen for cognitive impairment, complete tests in 15 minutes, or provide a longer screening battery. Are you trying to conduct multiple repeat tests in a short period of time? It's really too much of a complicated issue with too many potential uses of these tests to boil it down to a which test is best question.

But since we're a DoD organization, I do want to point out that the ANAM4, as the DoD's primary NCAT, I think stacks up reasonably well against the others. In other words, there's really no clear evidence that the ANAM4 – that using that is a big mistake on the DoD's part. I think it stacks up really well against the others.

Also, what's interesting is the type of analysis utilized can paint a really different picture. Just look at our results. ImPACTs did not do well when comparing controls to those with mTBIs; however, it did do very well when identifying cognitive impairment as we defined it. We all know the saying: There are lies, damn lies and statistics. So it's important to consider these statistics and, again, the intended purpose of the tests when evaluating the utility of NCATs.

The bottom line is we feel that there's a lot more work to do, and currently NCATs are still best suited as a screening tool or one component of a larger battery.

On to slide 85.

What we've really come to realize is that these head to head comparisons are somewhat of an apples to oranges comparison; or, as our moderator Dr. Gregory recently put it, a "messy to messy comparison," which I might like that better. What do I mean by that?

While each NCAT on the surface sounds similar, it's a computerized neurocognitive assessment tool. We're looking at things like memory and attention and speed, et cetera, on a computer platform. However, each NCAT has different stimuli; there's a different manner of presenting those stimuli; and there are different response methods. Think about CogState. It's a playing card motif; it looks very different than the others.

The same cognitive domain can also be measured in very different ways. And that influences direct comparisons. We see evidence of the NCATs referring to similar tests as different things, such as one referring to a particular test as processing speed and another NCAT referring to a very similar test as attention. And some of the NCATs report subtest scores, whereas others report composite or index scores. And those composite or index scores combine several subtests, so they might be more robust measures. So when you're comparing a subtest score to a composite score, you're not really getting perhaps the best comparison.

Also, each NCAP defines cognitive impairment in different ways; and none of those were entirely consistent with how we defined it on traditional tests. In some ways, it's kind of remarkable we found relatively similar results across the four NCATs in those analyses.

Participant effort was also defined differently on each test, again affecting who was included and who was dropped. We're doing so in somewhat different ways across the tests.

I like this quote by Kaminski and colleagues that is on slide 86 because I think it sums up the challenge of head to head studies in neurocognitive testing: "...test batteries are in fact measuring very different and unique characteristic traits of neurocognitive functioning... and ...not all neuropsychological test batteries are created equal."

So to think we can dump four into analyses and get clean results is a bit naïve, as we quickly learned.

Slide 87.

Where to now? Well, for our data, we plan to take a slow and methodical deep dive into the data and really into the issue of the apple to oranges or messy to messy comparison. We want to clarify what factors the NCAT scores are loading on and how those relate to traditional tests. This is going to necessitate using principal components analysis or other similar approaches to accomplish that goal. This is going to set the stage, I think, for more accurate direct comparisons.

We'll also look at alternative ways of investigating reliability – perhaps to include additional approaches to reliable change indices or regression-based measures. I said that we did some reliable change indices. We didn't find that the way we calculated them was particularly meaningful results and that's reviewed in our 2013 paper. I think there are some different ways that we can look at that that will be a little bit more meaningful.

From that we'll hopefully be able to determine if the reliability is more robust than we may believe it is. Or perhaps we'll further investigate if the current psychometric standards we use for our traditional tests is appropriate to use for computerized tests.

We're going to use various approaches, such as revisiting the raw data and using base rate analyses to assist with putting all the NCATs on the same playing field. So our comparison will be a little bit closer to apples to apples.

In general, the field of NCATs needs to take similar approaches to investigate the utility of these tests. We need to determine if baseline testing is useful and beneficial, or if normative comparisons are sufficient or possibly better for making return-to-play or return-to-duty decisions.

And I think overall, we need to be thinking about alternative methods, some of which I've described, for investigating the psychometrics. That is, we may need to break away from some of the standard approaches used with additional tests. NCATs are different animals in many ways, and that necessitates thinking about them in a different manner.

It was at this point there was originally an acknowledgement slide, but that was removed during the final draft. But I still wanted to take a moment to acknowledge those who contributed to this work. All this data was collected with a relatively small crew. Jacques has been a key component of the study. Another individual, Mary Alice Dale has been with this study from its inception; and she and Jacques accounted for the majority of the data collection.

We also had support from some other research and IT staff, to include Wes McGee, Katie Toll, and Alex Elliott. And thanks to Liz Dennison and Angelic Aarons here at our site for their help with the list searches. At DVBIC Headquarters, Karen Schwab, Brian Ivins and Felicia Qashu have been intricately involved with this study.

And (inaudible) was that third party statistical company that was very helpful in the initial analyses. We also had a lot of support from an IT company called A3 ITS. They helped us develop a platform that

allowed us to administer questionnaires and then seamlessly transition into the NCATs in a way that really made data collection very easy.

And finally, the four NCAT companies that participated in this, they provided us a lot of support and feedback; and we're very grateful for all of that.

We want to thank the Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury and the Defense and Veterans Brain Injury Center for inviting us to speak and for their support in bringing this webinar to you.

We sincerely thank all of you for your attention and participation in the webinar. We hope you found it valuable. The remaining slides contain references relevant to the information presented in this webinar. And now, with the time remaining, which I believe is at 10 or 15 minutes, we're going to take questions. If we don't get to your question or you have a burning question that comes to you as you ponder our talk before drifting off to sleep tonight, you can feel free to reach out to us outside of this webinar. We're happy to take any questions – again, our thank you.

Thank you, Dr. Cole and Mr. Arrieux – excellent presentation.

And, as Wes said, if you have questions for the presenters, please submit them now via the Q&A pod located on the screen. We already have a few questions from the audience.

The first question, excellent question: "If reliability is a necessary condition of validity and the reliability of these measures is adequate at best, can we expect them to be valid?"

Yes, that's a big sticking point. I think the easy answer is probably, no, not really. And perhaps that's why we see these sort of marginal effect sizes in some of our analogies. But I think it's a much more complicated issue. So I suspect – and others have stated this as well, so I'm not saying anything new here – that NCATs are looking at different cognitive abilities, or at least looking at cognitive abilities in different ways, than traditional tests do. So it could be that we don't really have a true gold standard to compare them to at this point.

Also, with regard to reliability, I'm not sure it's clear they're not reliable. It just seems to be at this point because reliability, as we've looked at it, is less than desired. I think things like alternative statistical approaches and additional considerations to the sources of error that's introduced in computerized testing, I think we might find a different story with regard to reliability. And so hopefully we can feel a little bit more confident about validity moving forward.

Agreed, and I have a follow-up question that kind of speaks to that a little bit. The question from the audience member is: "While the reliability (inaudible) on these tests is often disappointing, as you've spoken to, we've seen equally modest test-retest reliability on the traditional tests, such as the Rey-Osterrieth's figure test. Any thoughts on the same issues in traditional testing?"

Yeah, I mean, a lot of the tests have psychometric properties published; and they're usually pretty solid, but yet not perfect. And that's definitely a consideration. I think that speaks to a larger problem with some of these tests that we have, and I think that's something that's very important for us to consider when we are looking at all psychometric properties, especially things like validity, that have reliability making up a critical part of that.

So I don't think that it's necessarily anything that's unique to computerized tests. I think it's something that's been facing the field of neuropsychological testing for a while. But I feel like computerized tests just present some additional challenges that maybe will force us to think about these things in different ways than we think about them in traditional neuropsychological testing.

Thank you for that. We have more questions. Here an audience member asks: "Regarding the order of administration, do you think there are clinical situations in which a patient would be given more than one

NCAT in the same sitting." Because as you know, in the presentation you spoke to potential differences in augur effects.

Off the top of my head, I can't think of that necessarily happening. I think that's going to be more applicable to research studies. The only way that I could think of that happening is if we ended up finding out that certain pieces of -- I'm just going to use some examples. A few subsets of CNS were really good at looking at memory, whereas CogState was great at reaction time and ANAM was great at attention. And you kind of built a battery around subtests from those computerized measures.

But otherwise, yes, I think that's a great point. Clinically I'm not sure I would see a reason to do that, at least not yet. But for research purposes, as we're pitting these tests up against each other, I think it's an important consideration.

Thanks. Another question: "How are the instructions administered to the patients in these tests?"

We had an instruction screen that came up before each different kind of phase of the data collection. And then we also had a test proctor available to answer any questions. We tried to administer each NCAT as similarly as possible to the way that they instructed them to be administered. So we tried to be as standardized as possible.

And now on to a question about effort. You've had a number of questions about the effort scores with some of the service members. And one from the audience: "Why were the individuals with poor effort removed from the study? That seems to be a prevalent concern with administering the ANAM to service members for pre-deployment assessment."

Well, we removed them because their data would throw off the rest of the data. We wanted to look at data that was as clean as possible for the sake of establishing the psychometric properties. I certainly think participant effort and looking at the nature of how people who are putting forth poor effort and maybe some of the reasons they're doing that, I think that's absolutely a very valuable and important thing to investigate. But I think just because of the goals of what we were looking at, we felt it was best at that stage to remove them from the analyses because their data was really skewing the results.

Great, and a question from the VA perspective: "In your sample, the subjects were generally tested within 72 hours of injury. Can you speculate on the use of NCATs, and ANAM in particular, to predict impairment relative to traditional paper and pencil exams, let's say, in folks whose mTBI occurred months or years ago, which is what is commonly observed in VA settings."

Yeah, I would say the current clinical recommendations in the literature would probably suggest that they're not indicated for that use. If they're going to be used, then you may want to have a baseline test from their active duty days or use them as a screening tool only with a big asterisk beside them. I would say these are best suited currently for a screening tool as part of a larger battery of assessment during the acute phase of brain injuries.

And I'll end with one last question; it's an excellent question and something to think about in doing anything with NCATs: "Is it possible that NCATs are more sensitive to subtle effects than TBI and, therefore, comparing it to traditional tests is somewhat difficult? But on the other hand, what is the alternative?"

If I could answer that, I think I would solve a lot of our problems. I think, yeah, absolutely, they could be more sensitive because what we've seen time and time again is that these reaction time measures seem to be the most sensitive. There's some really interesting research that's out there on that and that's emerging on the usefulness of reaction time measures. And these tests are potentially much more sensitive at measuring reaction time than our traditional paper and pencil tests. So they could be -- especially during that acute phase -- much more sensitive to that. And that could have an impact on the psychometric property.

I think the issue of the lack of a true gold standard is a problem, and it makes this a really big challenge. And that's one of the things that I really was trying to speak to, that we need to take some alternative approaches to this and consider some different ways statistically of looking at these because, yes, I think this is an issue. And I think it's a big challenge that we have.

Thanks, Wes. That was a hard question but a beautiful response.

So on that, we'll begin the close. As you heard, Dr. Cole and Mr. Arrieux did a really nice job of writing a rich overview of the literature of NCATs and their psychometric properties and then described very interesting and (inaudible) incredibly valuable findings from their study conducted at Bragg comparing four NCATs.

Thank you both for presenting today.

For the audience, after the webinar, please visit www.dcoe.cds.pesgce.com to complete the Online CE Evaluation and download or print your CE Certificate for attendance. The online CE Evaluation will be open through December 24, 2015.

And to help us improve future webinars, we encourage you to complete the feedback tool that will open in a separate browser on your computer. To access the presentation and resource list for this webinar, you may download them from the File pod on the screen or on the DVBIC website. An audio recording and edited transcript of the closed captioning will be posted to that link in approximately one week.

The Chat function will remain open for an additional 10 minutes after the conclusion of the webinar to permit attendees to continue to network with one another.

The next DCoE psychological health webinar topic is Year in Review, Clinical Practice Guideline: 2016 Post-traumatic Stress Disorder; and that is schedule for January 28, 2016, from 1:00 p.m. to 2:30 p.m. Eastern Time. And the next DCoE TBI webinar topic is entitled Do Head Injuries Cause Chronic Traumatic Encephalopathy, and is schedules January 14, 2016 from 1:00 p.m. to 2:30 p.m. Eastern Time.

Thank you all again for attending today and have a great rest of the day.