47| Uniform Test Score Labeling – With Dr. Tom Guilmette

June 1, 2020

NAVIGATING NEUROPSYCHOLOGY

This is an audio transcription of an episode on the Navigating Neuropsychology podcast. Visit www.NavNeuro.com for the show notes or to listen to the audio. It is also available on the following platforms:

Listen on Apple Podcasts | Listen on Google Podcasts | LISTEN ON STITCHER | Listen on Spotify

Speakers: Tom Guilmette, Ryan Van Patten, John Bellone

Intro Music 00:00

Ryan Van Patten 00:17
Welcome, everyone, to Navigating Neuropsychology: A voyage into the depths of the brain and behavior, brought to you by INS. I'm Ryan Van Patten...

John Bellone 00:26

...and I'm John Bellone. In case you didn't hear our announcement during the last episode, we are now proud partners of INS, the International Neuropsychological Society. One aspect of the partnership is that NavNeuro episodes now qualify for continuing education credits. These are APA approved CEs that you can get by going to navneuro.com/INS.

Today we speak with Dr. Tom Guilmette about uniform test score labeling guidelines in neuropsychology. Tom is a board certified neuropsychologist and is affiliated with Providence College and Brown University.

Ryan Van Patten 01:03

The context for this episode is a project that Tom has been working on with multiple collaborators for a number of years, attempting to establish a system of qualitative descriptors that can be used by all neuropsychologists. This has come to fruition recently with a publication from the American Academy of Clinical Neuropsychology, or AACN, in their journal, The Clinical Neuropsychologist.

John Bellone 01:28

You may have seen this discussed on the neuropsych listservs and may have already begun incorporating the new system into your reports. We're going to link to the paper and the new score labels on our website at navneuro.com.

Ryan Van Patten 01:40

This topic is especially near and dear to our hearts because one of the primary goals of NavNeuro is to bring consensus and uniformity to topics such as this within our field. John and I both feel strongly that the new test score labeling system is well designed [and] clinically useful. We really encourage all neuropsychologists to give careful consideration to implementing the system in their clinical practice as we have both done. And, with that, we give you Tom Guilmette.

Transition Music 02:09

Ryan Van Patten 02:19

Alright, we're here with Tom. Welcome to NavNeuro.

Tom Guilmette 02:22
Well, thank you very much. Thanks for having me.

Ryan Van Patten 02:24
Yeah, our pleasure. So we can start by asking you to give us a brief background on the concept of establishing a uniform test score labeling system.

Tom Guilmette 02:33
Oh, wow, I'm not sure that there is a brief background.

John Bellone 02:36
[laughs]

Ryan Van Patten 02:37
[laughs]

Tom Guilmette 02:37
There's a long and labored background. I can give you that version.

Ryan Van Patten 02:40
Fair enough.

Tom Guilmette 02:43
I think anyone who has been trained as a psychologist, or especially in clinical neuropsychology, has realized that as part of one's own training, as you go from supervisor to supervisor, you have to adopt their system of how they label test scores. I was aware of that in my own training and then when I had trainees, I was aware that they would come to me and they would say, "Well, Dr. Jones used to use this label for this score. How come you don't do that?" Well, they weren't quite that abrupt.

Ryan Van Patten 03:09
[laughs]

Tom Guilmette 03:11
But they had questions. And they were legitimate, of course. My own experience, my own training, the training of others that I trained, as well as getting reports from

folks where they didn't actually - let's say I was doing a reevaluation on a patient and the patient had a previous neuropsychological evaluation, and the report would come to me, for example. And let's say if there were no test scores actually listed, only a description of the test score, like "This score was impaired" or "Very low, poor, whatever", I would have no idea what that meant. It was frustrating as a clinician, frustrating as a trainee, as a supervisor.

So then in 2008, Leigh Hagan and Tony Giuliano and I decided to test this question empirically. We developed a survey that we sent to board certified neuropsychologists with a little scenario about someone who had sustained a brain injury. We gave them, the survey respondents, a list of many different standard scores and asked them to provide a label for each standard score. We found that in the upper end of the distribution, the labels were relatively consistent, although not fully. But as the scores got lower, the number of different descriptors just expanded. I think for some of the lowest scores, there were up to 25 different labels used to describe the same standard score.

John Bellone 04:34
Wow.

Tom Guilmette 04:34
So we then had clear empirical evidence that this was an issue in the field, for lots of different reasons - for trainees, for referral sources, certainly in a litigation context or in a courtroom in a forensic assessment or testimony. It would really leave someone very vulnerable to the fact that we don't, as a field, we didn't then have any uniform way of describing test scores. Yet, the other thing we found out was that qualitative descriptors are the most common labels that people use to describe test scores. So in their own reports, in non-forensic reports as well as forensic reports, the use of “average”, “high average”, “low average”, “below average”, “impaired” was the most common way of describing test scores. So clearly, they were used a lot, but they were used in an inconsistent and non-uniform way. So that was a problem for the field. And we punted in our discussion [and] basically suggested that some professional organization in psychology should write some sort of position paper on this and try to give us some sort of clarity and uniformity. That was in 2008.

John Bellone 05:39
Gotcha. You maybe didn't know at that time that it was going to be you who was going to write that paper. [laughs]

Ryan Van Patten 05:43
[laughs]

Tom Guilmette 05:43
I had no idea. Had I known that I probably wouldn't have written that sentence.

John Bellone 05:46
[laughs]

Ryan Van Patten 05:47
You tried to punt [it] but it actually came right back to you.

Tom Guilmette 05:50
It did, right. Well, the wind was very strong, in my face, so it just kept on coming back.

Ryan Van Patten 05:54
[laughs]

John Bellone 05:55
So the goal of the AACN Consensus Conference was to determine a labeling system that would be uniformly implemented. That means that, ideally, all neuropsychologists would adopt this system, but that it's not mandatory. Is that right?

Tom Guilmette 06:11
Correct. Yes. That's exactly right. This is a recommendation. It's not a standard. So unlike the APA ethical standards, where that's a list of do's and don'ts, this is not that way. We consider this to be sort of best practices. But it is strictly voluntary and any individual clinician is free to reject the whole thing, or take what they like, or leave what they don't like behind. So it's strictly voluntary.

John Bellone 06:38
The system is meant to be used in place of those provided within specific test manuals, right? For example, even if the Wechsler label is “extremely low” for an IQ score, you wouldn't necessarily use that label. Would you apply the consensus statements label to the percentile?

Tom Guilmette 06:57

Yes. So one of the problems that I suppose I should have mentioned when we were talking about the nature of the problem, or the history of wanting to establish a uniform test score labeling system, was that clinicians would be looking at a manual for a particular test and look at the publisher's recommendation for what to call a standard score of X and then they would look in another manual, and if the patient or client sustained the exact same standard score on a different test, that publisher had a different recommendation. So in a report, a clinician might be calling a standard score of 85 as "low" according to one test publisher, "very low" according to another, "mildly impaired" according to another. And the clinician was really caught in a bind because he or she is trying to stay true to the test publisher's recommendation but at the same time being quite inconsistent in his or her own report. Calling the exact same standard score a different qualitative descriptor based upon the different descriptors given in the manuals.

Ryan Van Patten 08:09

I like your recommendation to use the same system, even if it's not what's recommended by a test publisher. I think within our reports, ideally, within our entire field, if possible, we have consistency. So I'm on board with that. In this vein, talk about a few of the other systems that are out there. And, briefly, the pros and cons of these systems - Wechsler, Heaton, etc.

Tom Guilmette 08:34

Right. So we looked at all of those systems, and we struggled. Looking at each of those, each one has their own advantages, disadvantages. When we were looking at, and especially looking at tests with a normal distribution, we wanted our labels to be non-judgmental, or as much as possible, and to really simply try to reflect a person's performance along the normal distribution. Where does this person's test score fall relative to other people who have taken this test?

One of the main problems - you mentioned Heaton and Wechsler - one of our concerns about the Heaton recommendations is that in that system, scores are labeled as “impaired” in some cases depending upon where it falls. And one thing that we easily came to consensus about as a group was that we strongly recommend that clinicians do not call any score “impaired” because a score cannot be impaired only an ability or a function can be impaired, not a test score. That the definition of impairment or the application of that term needs to be used within context, within the individual's history, background, referral question, etc. So to call

any score impaired, we thought was a misnomer and potentially misleading. So that was relatively easy for us.

The Wechsler system certainly has its advantages. And as you can see, we use a number of the Wechsler labels. Probably one of the most challenging things for us or the label that we did not particularly want to endorse was the “borderline” label. We struggled with knowing exactly what that meant. We thought it was ambiguous - borderline, what? Borderline impaired, borderline normal, borderline average? So we decided to eliminate that label. We looked at a number of different labels.

We established what we thought were the appropriate labels at the end of our consensus meeting, which was in June of 2018, the day before the AACN annual meeting. But those labels could change based on feedback that we got at the meeting later. And after we posted our suggestions online on the AACN listserv. And so we're continually getting feedback, people's suggestions, their criticism, what they did and didn't like about various labels. And so through that very long process, we ended up coming up with labels that we have now, which as I said, is consensus. Everyone, all 22 of us, agreed on these labels for both the normal distribution, non-normal distribution, and the definition of impairment. So, again, pros and cons with other systems, we tried to take what we liked the best out of all of them, and incorporated that with the feedback that we got from AACN membership.

Ryan Van Patten 11:22
So you had mentioned the issue of whether or not a test score is impaired. I'd like to just stay with that for a moment.

Tom Guilmette 11:30
Sure.

Ryan Van Patten 11:30
I think that's a big shift for many people who are used to calling test scores impaired. My understanding in reading papers [is] that you're advocating for calling an ability impaired but not a test score. Can you elaborate on that?

Tom Guilmette 11:45
Sure. Yeah. For some folks who have used a labeling system for decades, it's what they were trained with. For those folks who are accustomed to calling a certain standard score “impaired”, this would be a very big shift for them to abandon what

they're accustomed to using. Both you are relatively new in your careers, I'm not so much, and so I can tell you that old habits die pretty hard. What you're trained with is what you tend to rely on. That doesn't mean that we can't be flexible. But if you've been using a certain system literally for decades, then it's very hard to give that up. However, this was an issue that we came to a consensus on very easily. We simply thought that it was an inappropriate use of the term impaired to call a score, this intangible thing, impaired. No score is impaired. Only an ability or a function can be impaired and that's based upon the judgment of the neuropsychologist incorporating all of the data about the patient. To simply take any standard score, and to call this unit score impaired, does not take into consideration all those factors that the neuropsychologist would. Impairment really is an interpretation. It's the interpretation of the neuropsychologist to call a function or inability impaired. A score is not an interpretation. A score, again, from our perspective, is simply to denote or to identify the location of a patient's score along a distribution.

John Bellone 13:17
Right. We're not test bound or taking a concrete approach that each score has to have inherent meaning. It's just one bit of information that the neuropsychologist has to incorporate into the overall conceptualization. We'll get into your specific system in just a second, but that's also my understanding of why you included the actual word “score” after each label.

Tom Guilmette 13:42
Yes. Right.

John Bellone 13:43
Just to emphasize that point.

Tom Guilmette 13:45

That's exactly right. Yes. We want folks to know that we are calling this score this thing, and it pertains only to the score, not to the ability or the function. Exactly. Right.

John Bellone 13:58

Great. Before we get into more of the weeds, just generally, I'm curious what the benefits of uniformity are? Some of them will be obvious to our listeners, but some of them might not be.

Tom Guilmette 14:10

Well, uniformity. I think it reflects a maturation of our field, frankly. Rather than having multiple disparate systems floating around up there based upon the whim - is not the proper word, but based upon the decision of any particular clinician. For us to communicate with each other colleague to colleague. For trainees to discuss cases with supervisors. Our ability to be uniform in our description of test scores to our referral sources, both within psychology and/or neuropsychology or outside of the field. Certainly within a forensic context, it's necessary for us to be consistent in how we describe score. I think it really reflects a maturation of our field where we are trying to pull folks together and to get greater consensus within our profession, that we all should be using the same labels. We don't necessarily all need to come to the same conclusion about what a test score ultimately means relative to a person's diagnosis or not. But what we simply call a score along a distribution of scores from a normative sample seems, to me, to be pretty basic in terms of how we should be communicating these scores to each other within the profession and also outside of the profession.

John Bellone 15:29

Yeah, I completely agree. And, like you, I had seen many different systems across my supervisors and training sites. It was always so confusing to me that we didn't have just one uniform labeling system.

Ryan Van Patten 15:47

I agree. Sometimes my experience was that it was communicated to me that the system I was using was wrong and a new system was right.

Tom Guilmette 15:56

What? [laughs]

Ryan Van Patten 15:57

And that was confusing to me. Why would one arbitrary system be right and another be wrong? So then everything that I've learned is wrong? That doesn't make sense. All that to say, I completely concur with what you're saying, Tom.

Tom Guilmette 16:10

Yeah. Both your points are exceptionally well made. It's so confusing for trainees to go from Dr. A to Dr. B. And, again, you know, Dr. A calls a standard score of 75, whatever you might call it, and Dr. B calls it something different. As a trainee, you're

really caught in the middle. So now you have to shift your paradigm every time you move from one supervisor to the next. And really the labels, the words that we use, are somewhat arbitrary.

Ryan Van Patten 16:41
Yeah.

Tom Guilmette 16:41
We felt strongly that we still needed to agree on a label. We can arbitrarily assign labels. I mean, not completely arbitrary, but, you know, we can talk about the statistical factors or properties of a score, a percentile rank, the z-score, where it falls relative to the mean, etc. So we're not just whistling in the dark, but the words that we use, we need to decide upon those for all the reasons that we have mentioned here today.

Ryan Van Patten 17:15
John and I have had this conversation offline and we've said something very similar. You know, I'm very flexible in what system ends up being implemented, where we draw the line between “below average” and “average”, etc. I could go conservative or liberal or somewhere in the middle - as long as we all agree. [laughs]

John Bellone 17:33
That's the most important.

Tom Guilmette 17:35
Right. Yes, exactly.

Ryan Van Patten 17:36
So drilling down, a specific quote from your article is that, "deviations may render clinical interpretations vulnerable to attack in litigious contexts." What did you mean by that?

Tom Guilmette 17:51
We meant that in the absence of uniformity, and in the absence of an agreed upon method of describing a standard score with a certain qualitative label, then in cases where the outcome of a decision or clinical report or a score has relevance to, let's say, the outcome of civil litigation, or in a criminal case, or in cases of someone

receiving special education services, that the difficulty for the clinician in calling something different from the person sitting next to them who's also a clinician, but calling the same score something different, or using two different labels for the same standard score because they come from two different test manuals or from two different tests and the manuals describe the scores differently, leave someone vulnerable to the attack of, "So which is it doctor? Is this score poor? Is this score impaired? Is this score very low? What is it, doctor?" You can see how that would cause significant problems within a litigious or forensic context because the lack of uniformity makes us look non-committal and that we can't agree as a profession, what to call this thing. So why should I believe this individual clinician, when the whole profession can't agree on what to call this label?

John Bellone 19:12
Right, on cross examination the attorney is going to ask, "Well, you know, Dr. Guilmette, Dr. Martinez who has more publications than you does it this way." [laughs]

Tom Guilmette 19:21
Right. Exactly. Yes.

John Bellone 19:22
"Clearly, you don't know what you're doing." Right. Exactly. [laughs]

Tom Guilmette 19:24
Right. Exactly.

Ryan Van Patten 19:25
It weakens our field as a whole.

Tom Guilmette 19:26
Yes, for sure it does.

Ryan Van Patten 19:28
What would you say were the factors or the barriers to creating a uniform system? Why wasn't this done before 2020?

Tom Guilmette 19:36

That's a great question. Why wasn't it done before 2020? [laughs] I'm not sure. Maybe because it's hard to do. [laughs] I think maybe because people were comfortable with their own scoring system. Perhaps they thought their scoring system was correct, so why should they give that up? Or maybe they thought, “People should adopt my system.” It was striking when we went to the AACN Board of Directors in 2017 and Jerry Sweet and I proposed this consensus conference, nearly everyone at the table was shaking their heads vigorously, "Yes, yes, yes. We need this. We need this. We need this." And, yet, no one had really acted on this much before.

We tried this in 2017 but not nearly in the same way with a consensus conference. But we tried to do this more informally via email and that kind of thing. We didn't make a whole lot of progress, I think because it was hard to do. Perhaps because people didn't think that there would be acceptance within the field or that the field was not ready for this. That's a really good question. You know, the timing was, I guess, finally right. I'm not sure why but I'm glad it was ultimately.

John Bellone 20:51

Right. Ryan and I have talked about the problem of inertia a lot…

Tom Guilmette 20:54

Yes, for sure. Right.

John Bellone 20:56

…with a number of different aspects, different technologies. And this is along those lines too. People are doing what they're comfortable with and it's hard to change. We understand that.

Tom Guilmette 21:07

Right, yes.

John Bellone 21:08

Are there any downsides to having a uniform system? One that I thought of as a possibility is that is it a problem to have the same system for diverse settings and populations?

Tom Guilmette 21:19

Sure, yes, that's a potential downside to this. And therefore I think clinicians will need to make their own judgment about whether or not this system applies best to their populations, to their settings, the context that they see patients in. So that's always a possibility. Our view is that we still are advocating for folks to use the system. But clearly, again, if clinicians feel that it's not right for them or for their setting, then they certainly do not have to use this system. I would also add that we don't see these labels as being etched in stone. I really view this as being the beginning of a process. Perhaps there will be another consensus conference in 5 years, 8 years, 10 years, after folks have used this system for a while and our thinking becomes clearer about it. Then perhaps these are revisited at some point in time in the future. In fact, I would hope that they would be revisited at some point in the future. I'm not going to revisit them.

Ryan Van Patten 22:18

[laughs]

Tom Guilmette 22:18

But you all feel free to revisit them.

John Bellone 22:18

[laughs] You're done.

Ryan Van Patten 22:21

[laughs]

Tom Guilmette 22:23

So I really see this as a process. We learn something from things as we implement them, and hopefully it will be widely implemented. We use it for a while - years, maybe, and then we see how well it fits and how we can improve upon them. Perhaps those special circumstances or context or population that folks might deal with, then perhaps there are different labeling systems to use within those populations. I don't know how all this plays out eventually. But I don't see our consensus statement as being the final chapter in this process.

John Bellone 23:02

Let's talk more specifically about the consensus conference system. Some might call it the Guilmette System. I don't know if that's… [laughs]

Tom Guilmette 23:11
[laughs]

Ryan Van Patten 23:11
We're coining that right here. [laughs]

Tom Guilmette 23:13
No, no, no. Please don't.

John Bellone 23:16
[laughs]

Tom Guilmette 23:16
Just to be quite fair, there were 22 people who were part of this, who met together in June of 2018. Brilliant folks. Nationally prominent folks - pediatric, adult, geographical distribution, gender, ethnicity. We really came together. This was truly a group effort. So we'll just call it the Consensus Conference Recommendations.

Ryan Van Patten 23:42
That's fair. [laughs]

John Bellone 23:42
Fair enough. Fair enough. [laughs] So let's start with tests that have a normal distribution first. The ranges are based on standard scores - so a mean of 100, standard deviation of 15 - rather than integer standard deviation cutoffs.

Tom Guilmette 23:58
Yes.

John Bellone 23:59
That's led to a seven category model, instead of the other one that you had considered which was a five category model. What was the rationale for the seven category [model]?

Tom Guilmette 24:09
There were really two. One was that clinicians were really used to the seven category model, so we thought that it would be more easily adopted by clinicians.

While you could argue there are some potential benefits for using the five category model based on standard deviations, we also thought that, in addition to clinicians being more accustomed to the seven category model, there are also finer gradation of labels that we thought that clinicians would prefer anyway. So both because it's more commonly in use and because it gives clinicians a finer gradation of how to describe labels, we thought that the seven category model was more advantageous.

Ryan Van Patten 24:59
How did you decide on specific test labels: average, above average, below average, etc.?

Tom Guilmette 25:05
We looked at previous systems. At the end of the day in June of 2018, we had come to a consensus - again, so everyone agreed on these seven labels, which were ultimately changed. We had, for example, from 90 to 109 was average. And then 80 to 89 was below average. So we jumped right from average to below average, rather than low average, which is what we finally agreed upon in the end. We had all agreed on the set of labels based upon we felt they were the least judgmental, the best descriptive, incorporating the best of what was out there.

Then, three days later, we presented our labels in an open forum, basically like a town hall meeting at the AACN conference in 2018. We had like open mic night, in that folks got to ask us questions, our rationale, give us their suggestions, and give us their critiques. And there were a lot of challenges to what we had suggested. [laughs] So we left San Diego and we ended up posting our consensus recommendations onto the AACN listserv for about two weeks in July of 2018 and got more information, or more feedback, from clinicians who either did not attend the AACN conference or had but wanted to say more. So our consensus essentially crumbled from there. [laughs]

John Bellone 26:44
Oh… [laughs]

Tom Guilmette 26:46

Because all 22 of us who were reading these comments, we were all obviously at the open forum meeting on the Friday of the AACN annual meeting in San Diego. And the feedback that we got, or the message that we got, we felt that the original labels needed to be revisited because we had push back on a few issues. So we

went back to the drawing table in some ways. And over the course of - I actually don't remember how long it took us after that. It took us a long time. Nearly a year, I think. Through email and other conversations, discussions, sending it back to the subgroup that was looking specifically at labels for the normally distributed tests, and finally evolved. This actually was the very last thing that we were able to come to a consensus on of the three things that we're looking at. And that was, I think, nearly a year after we had presented our model in 2018 in San Diego to the AACN membership, and then posting it online on the AACN listserv. So there was a lot of rethinking. Other opinions emerged. We tried to find points of compromise, adding qualifiers like mildly, moderately, etc. And so we finally were able to come up with the consensus that we did. But it was with a lot of deliberation and a lot of thought about it.

Ryan Van Patten 28:13
Another more specific decision you ended up making regarding the labels, I'm referring to the tails of the distribution, you were deciding between “extremely” and “exceptionally” low or high.

Tom Guilmette 28:25
Yes.

Ryan Van Patten 28:25
Can you briefly talk about the conversation there?

Tom Guilmette 28:28
Yes. That was more a semantic argument or a semantic concern. There were some folks who thought that “extremely” did not really capture the rarity of those scores. And there was not a strong feeling within our group, about the use of the word “exceptionally” versus “extremely”. But outside of our group, there seemed to be a stronger sentiment about using “exceptionally” versus “extremely”. So we decided to go with “exceptionally” because that seemed to be more of what the general AACN group preferred rather than extremely. I'm sure there were some folks who had no opinion about this whatsoever, but those that did seem to side on the “exceptionally” versus “extremely”. So we went with that.

Ryan Van Patten 29:18
This is just a window for listeners into how detail oriented you ended up getting. The semantics between words like “extremely” and “exceptionally” were a point of debate.

Tom Guilmette 29:28
Yes, right.

John Bellone 29:29
Right. And other modifiers, “moderately” or “mildly”, and then whether to say “impaired” or not, and low or high and all that. You really went through all those decision points.

Tom Guilmette 29:38
Yes, yes. The most contentious range of scores was 80 to 89, which we had originally called “below average” and we got a lot of pushback on that. We revisited that and then some people felt that to call a percentile rank of 24 or 23 below average was too stringent. We then talked about perhaps calling the 80 to 89 range “mildly below average”, but then we were stuck with, "What does “mildly” mean?" "Slightly”? But what does “slightly” mean?

Ryan Van Patten 30:15
[laughs]

John Bellone 30:15
What a nightmare.

Ryan Van Patten 30:16
So many adverbs and adjectives.

Tom Guilmette 30:17
Right? Yes. We floated all of those among all 22 of us. And, actually, when you look at the 80 to 89 label, many of the extant labeling systems use “low average”. So we went back to the Wechsler label and we thought that that was least contentious ultimately. Everyone is still not crazy about that I'm sure out in the field. But we could come to a consensus about “low average”, in part, because it seemed clearest among the options. It didn't have a modifier and it's what most clinicians

are used to using in that range anyway. So we thought it would be the most accepted.

John Bellone 30:59
You're never going to make everybody happy or find the perfect factor.

Tom Guilmette 31:03
Exactly. Right. Yes.

John Bellone 31:06
To get down into the nitty gritty details a little bit. What do you do in the rare instances where higher standardized scores means worse performance? I'm thinking of something like the Hooper Visual Organization Test, for example.

Tom Guilmette 31:19
Right. Yeah.

John Bellone 31:20
You choose the label that reflects the distinction, I'm assuming, right?

Tom Guilmette 31:23
Right. Yes. I think what I would suggest is to - and we didn't address this specifically, I don't think, in the actual paper. But what I would suggest is to transform the scores as if, you know, so you transform a z-score or a T-score from above the mean to below the mean, and then use that label that you would use to describe the score.

John Bellone 31:44
Just flip it. Okay.

Ryan Van Patten 31:45
So poor performance should always mean low average, below average, exceptionally low, etc.

Tom Guilmette 31:51
I think that's the least ambiguous.

Ryan Van Patten 31:53
Yeah, I agree.

John Bellone 31:54
Then the other detail that I thought of, because I've been adopting it and so these questions come up as I'm using the system, what should we do when a percentile range spans two different labels? So, for example, I've come across 6th to 10th percentile on the BVMT or the Wisconsin. I've been putting "low/below average." What do you recommend?

Tom Guilmette 32:17
I think that makes perfect sense to me. Yes, sure.

John Bellone 32:19
Okay.

Tom Guilmette 32:19

That sounds reasonable. That's keeping within both the spirit and the pragmatics of what we were trying to recommend. So I think that makes perfectly good sense to me.

Ryan Van Patten 32:30

A very important point here would be scores - z-scores, percentiles - that fall near a cut point and what do we do there? So, in your paper, you mentioned that, "clinicians should give careful consideration to labeling scores near cut points, including consideration of the error band." How does being near a cut point alter how you label the score? For example, if a score is at the 8th percentile? Do you choose between a label of low average and below average? Or do you go with what is suggested but modify your interpretation?

Tom Guilmette 33:06

Well, ultimately, the interpretation is the interpretation based upon all the factors that you would use in deciding on what test scores actually ultimately means. But I think here, clinicians would be free to do what makes the most sense to them. What I would say is that if you're comfortable calling something "below average", rather than "low average/below average" - John's comment, I think, is germane here about using when a percentile rank falls between two ranges, to perhaps use both. So I think that makes some sense. If you're right on, let's say, a standard score of 80,

the error band or standard error could obviously mean that your score is actually in the upper 70s rather than 80. So that is below average, rather than just low average. So I think depending upon how strict you wish to be, you can either say, perhaps as a disclaimer on your table of report or how you report the scores, that, "All scores contain some measure of error. I am reporting scores here as they fall or as they were obtained by the individual, but recognize that there might be some variability ultimately if we were to calculate the center of measure for any of these scores." So you could, I think, either have a blanket statement advising the reader that while you're calling the score by a certain label, there's some error within that score. And so actually if it's close, then it could dip into another label category. Or you could, as you describe any individual score, you could call it, again, "low average" or "low average to below average". I think that would be up to the clinician to decide whatever the most comfortable with.

Ryan Van Patten 35:00

Yeah, that makes sense. Of course, making our interpretations, we should not be wedded to qualitative descriptors. We should look at the standardized scores, the percentiles, and use our pattern analysis abilities, as opposed to feeling like we're inherently tied to what a qualitative descriptor says, I think. Along those lines, the fact that you remove the word "impaired" from the qualitative descriptors helps because, in the past, when we would call a score "impaired" the clinician might feel like because this one score was impaired among all other CVLT-2 scores, well that means memory is paired.

Tom Guilmette 35:40

Right. Yes.

Ryan Van Patten 35:42

Part of what we're getting by removing the word "impaired" from test score labels, but even more broadly, I think, we should not be so wedded to the labels and look at other metrics of performance.

Tom Guilmette 35:56

Yes, well said. Yes, I agree. If you're bound by a certain system that calls a score "impaired", then you're really handcuffing the clinician. So then you are now calling this impaired because you are bound by this system. So then what do you do in the interpretation? Do you ignore the word impaired? In my opinion and the opinion of the consensus conference, as I mentioned before, is that impairment is an interpretation. These labels are only meant to describe or define where someone's

performance falls along a distribution. What you do with that is the clinician's discretion. It's the clinicians interpretation that determines what to do with that score. You are not bound by a label. The clinical interpretation is up to the individual professional based upon his or her experience, as well as all the other factors that one incorporates in arriving at a diagnosis or at a conclusion.

John Bellone 36:55

Along these lines, it seems like these labels are relatively conservative compared to the model that many neuropsychologists were previously using. For example, in the Heaton framework, “low average” goes down to the 9th percentile. Were these labels intentionally set to be conservative?

Tom Guilmette 37:16

Yes. I think, in part, because that was the feedback that we received. We wanted clinicians to have as much flexibility as possible. So the issue of calling the 80 to 89 standard score “below average” versus “low average”, you then start getting into the problem in some ways of, again, somewhat handcuffing a clinician who would not conceptually consider a score of 86 or 85, certainly at 86 to 87 still within the same deviation from the mean, as calling that “below average”. Many clinicians will call that “average”. We tried to strike a balance. I would say that these are generally conservative and are, I think, keeping with, again, what other scoring systems have used, generally speaking. This is not brand spanking new. Our labels are not coming out of left field - at least I hope that they don't seem that way. You know, these are well within the ballpark of what other labeling systems have used with the exception of getting rid of the impairment label. I think we tried to keep these generally as consistent as we could, based on other systems and ones that would be most acceptable to clinicians.

John Bellone 37:17

You mentioned the range of 80 to 89 being particularly a source of conflict. I can see the reason why and it's because that's when people - so my old system, which I inherited from my supervisors, was to call a standard score of 80 “mildly impaired”. And so now it's “low average”. It does frame my interpretation somewhat because I think about it in terms of z-scores. That's like a score of -1.2 or 1.3 maybe.

Tom Guilmette 39:07

Right. Yeah.

John Bellone 39:08
Or that’s “zed” for our non-American listeners.

Tom Guilmette 39:12
[laughs]

John Bellone 39:12
Zed scores. I just recently learned about that. In Canada, I think, they say zed scores instead. [laughs] So it does sort of frame my interpretation somewhat. I know it shouldn't. We shouldn't hang our diagnosis on those labels. But it does affect how we think about it somewhat. We're less likely to call a score that's labeled “low average” problematic than we would score labeled “mildly impaired”.

Tom Guilmette 39:37
I hadn't thought of that as being a potential outcome. If it is, I would actually welcome that. In part because I think that - I don't want to get too far afield here, but I think, as a profession, we tend to err on the side of overpathologizing and of not always keeping in mind that range of normal variability among folks, among neuro-healthy people. And so if it's - as I said, I don't remember this ever coming up in any discussion that we had about that particular label or its effect on anyone's interpretation. One thing that we tried to make very clear was that these labels should not actually - these labels are not interpretive.

John Bellone 40:20
Right. No judgments.

Tom Guilmette 40:22
They are descriptive, right? That's all. The interpretive part comes in another part of your report by a clinician. I hadn't even thought about that as a possible effect of these sorts of labels. But from my way of thinking, I don't think that's necessarily a bad thing. But that's me.

John Bellone 40:39
Yeah, I agree with that. We tend to overpathologize and hang our hats on one or two scores that are low without considering the normal variability that's inherent in every test profile.

Tom Guilmette 40:52
Sure, right.

Ryan Van Patten 40:52
You're referencing work on base rates of cognitive impairment in healthy populations, which is a whole nother conversation. Grant Iverson and others have done great work in that area. I agree with the idea that, certainly at times, we potentially are prone to overpathologizing. So it's an interesting discussion.

Why don't we move into briefly discussing tests with non-normal distributions, for example the Boston Naming Test. These measures where most people score at or near a perfect score. So the scoring system for tests with non-normal distributions is overall very similar to that of tests with normal distributions, with the exception that you're not providing qualitative distinctions between different levels of intact performance because we don't have the sensitivity, the precision, to discriminate people at that level. Is that accurate? Anything more to say there?

Tom Guilmette 41:51
No, that's all very well put. Exactly. You have captured our thinking completely. Yes.

Ryan Van Patten 41:57
Perfect. Tell us about suggested labels for performance validity tests - valid, invalid, and indeterminate.

Tom Guilmette 42:05
Right. This also was a point of contention, arguably, in terms of feedback that we got that clinicians had most concerns about. The first was our labels for the normally distributed tests, especially the 80 to 89 standard score range. And then my impression was that the second area was what to call performance validity test scores. I actually can't remember offhand what our initial labels were, but we definitely got some feedback about that. What we tried to do there was to, again, describe a range so that a score falls within the invalid range and the indeterminate range. We thought that this still gave clinicians the opportunity, as they should, to interpret what this score range means. So this score on this PVT falls within the valid range, or falls within, let's say, within the invalid range, or that this is an invalid range score. That doesn't mean that the test is invalid. It simply means that the score fell within this range that we described as being invalid. We felt that this gave clinicians still the most flexibility in terms of their interpretation. Also for those folks who do a lot of forensic work, they felt that also - the feedback that we got from

many of them was that these labels were the most defensible. I can't recall what our original labels were for this, but when this was presented or when this was discussed among some of the more forensically-oriented neuropsychologists out in the field, the feedback that we got, generally speaking, was that these were the most acceptable labels out of all those that we had discussed previously.

Ryan Van Patten 43:54
Yeah, great.

John Bellone 43:55
I want to go back - we were talking about interpretation a little bit. I know we don't want to derail the conversation too much and we could talk about interpretation for an entire episode, but I do want to bring up the question of whether we even need qualitative descriptors at all. We should talk about that a little bit. If we, as clinicians, are the only ones who are qualified to interpret the data - you know, the neurologist who's referring, the patient isn't qualified to interpret the data, the patient clearly shouldn't be interpreting the data. Do we even need those descriptors? Or should we, as neuropsychologists, rely more on the standardized scores instead?

Tom Guilmette 44:34
I think you could make an argument that, in a report, the results section is simply a table of scores with z-scores and T-scores or standard scores, and then go right to the interpretation part. The fact is that that's not how we do business. That's not what clinicians do. As I think I had mentioned at the beginning, the most common way of describing scores is through qualitative descriptors by far - that's what clinicians do. That's what they use. It's hard for me now to think about how someone might interpret these labels, who is not within the profession. We try to make the labels very non-judgmental and reflect a performance, again, that follows along some distribution.

I suppose that there are some more sophisticated referral sources for whom - T-scores, z-scores may seem really easy to us because we've been using them, but I'm not sure how easy they are to interpret for folks outside of our fields actually. I guess that's an empirical question. So, for example, there's some agreement or disagreement around should we be using z-scores, or T-scores or percentile ranks as a way of also describing where someone's score falls along a distribution. I think that, in some instances, a more sophisticated referral source may want to know what this score actually means or where it falls, in which case, I actually think that a label of below average, low average, average, for some folks is actually more

meaningful than a z-score or a T-score. Ultimately, what you do with all of that is clearly up to the neuropsychologist and to the clinician. But I think you could make an argument that the only reason to put in scores is in case someone sees your report later and wants to do a follow up.

John Bellone 46:27
Right. A neuropsychologist sees it later.

Tom Guilmette 46:29
Right, yes. Right. Exactly.

John Bellone 46:30
That's my purpose for including it. Otherwise, I feel like everyone else should rely on my impression section.

Tom Guilmette 46:37
Yeah, sure. But if you look, for example, at - if you get an EMG report from a neurologist, you'll get the interpretation but you also see all the little graphs about what was found. And, presumably, that's not for us. [laughs] That's for other neurologists who might see this report later on. So I see value and I always suggest that people use or include raw scores into the reports, even if it's a table and either standard scores, percentiles, or z-scores or something like that. If you're going that far, then it seems to be reasonable to also include a descriptor. Our recommendation was not that we use descriptors, our consensus statement was that…

John Bellone 47:21
[laughs] If you do…

Tom Guilmette 47:22
Right. If you do use these…

John Bellone 47:25
I completely agree.

Ryan Van Patten 47:26
That's helpful. A brief follow up on the comment you made about what it would look like for a lay person to interpret our qualitative descriptors. One term that we haven't yet referenced, where I'm really glad you replaced the term, is “superior” and “very superior”.

Tom Guilmette 47:44
Right. Yes.

Ryan Van Patten 47:44
In the past, very high scores were labeled as “superior” and “very superior”. I don't think it takes a whole lot of analysis to imagine where that can lead people to go. “I'm not superior”, or “I am [superior]”. It's a very problematic word. I'm glad that we don't use that anymore.

Tom Guilmette 48:02
Right. Yes. It's loaded with all kinds of judgments. That was one of our main goals, to try to keep these terms as non-judgmental as possible. Again, more descriptive rather than interpretive.

John Bellone 48:18
So now that the statement has been out for a few months, how has the reception been?

Tom Guilmette 48:23
Well, I haven't had any death threats so far.

John Bellone 48:26
[laughs]

Ryan Van Patten 48:26
[laughs]

Tom Guilmette 48:28
No one has firebombed my car yet. I think it has been really positive, actually. I think people are happy that there is some guidance. That there is some place where we're saying, "Okay, here's what we are suggesting." Now, again, you can reject them if you want to, that's certainly well within your professional judgment to

make. But I think folks are glad that there is a system that is hopefully being adopted by more clinicians than not. I think everyone has felt frustrated with these very different labels, as both of you have experienced in your training, obviously. So I think folks really feel that this is a step forward in our professional development.

John Bellone 49:12

That's my sense as well. I was actually surprised that some of the people who I thought might have been against it or slow to adopt it were on board. It's been on the listservs, and I have had the same experience that I think people are really enjoying it.

Tom Guilmette 49:27

Yeah, I have been really quite happy with how this has turned out so far.

Ryan Van Patten 49:30

Yeah, same here. So, Tom, thanks for talking through this qualitative descriptor system. Before we let you go, we'd like to touch on the process of coming to a consensus in the consensus statement because that's something I think many people don't know much about. Maybe you can shed some light on that. So what does the process of coming to a consensus look like?

Tom Guilmette 49:54

Yes. So the way it happened - well, first of all, so consensus statements are pretty rare in our field. They're much more common in medicine. In neuropsychology, the only previous consensus statement that I'm aware of was the 2009 consensus conference on effort, response bias, and malingering. That was the only previous consensus statement or conference in our field, as far as I know. So the test score labeling consensus conference is only the second in our field. If you go into medicine, there are a lot of them.

In 2014, there had been other discussions about this problem on the AACN listserv, and Manny Greiffenstein had approached me about going to the AACN Board of Directors and seeing if we could ask or create a committee to look at this issue and to come up with a position statement. We tried to do [it] for two years and we had about 25 or 27 folks, and we tried to do all of this through email. We were not seeking consensus, we're basically looking at a majority. And progress was very, very slow going. In the summer of 2016, we presented what we had agreed to at that point - again, not a consensus so this was a very different sort of thing. We submitted a draft of what we had come to so far to the AACN Publications

Committee. Very sadly and tragically in the summer of 2016 Manny Greiffenstein passed away. So the committee sort of fumbled around for a while. We weren't sure what direction to take. So I approached Jerry Sweet at the time - I was co-chair with Manny of that 2014 group - I approached Jerry Sweet who was chair of the Publications Committee, and asked him what he thought about what direction we should go. He said that he thought this was important enough to approach the board of directors to create or to do this as a consensus conference. So we approached the board of directors with our model, which is basically the same type of model that was used in the 2009 consensus conference on malingering. And they said, "Yes, let's do it."

So then we went out and we generated a group of clinicians that we tried to balance by specialty and gender and ethnicity, and create three different subgroups for the normal distribution group, the non-normal, and the impairment label. [We] had readings that everyone did beforehand and then came to the conference and [I] kept my fingers crossed that we could come to a consensus. We did, but as I said, that consensus eroded after we got feedback. We were able to come back again because I think we all met together in one day. And so there was real cohesion in our group. I think that really facilitated the process of finally coming to a consensus on these labels.

Ryan Van Patten 52:48
What topics within neuropsychology would you say rise to the level of needing a consensus statement?

Tom Guilmette 52:55
Boy, that's a good question. I mean, there's no specific topic that comes to mind at the moment. You know, there have been a number of good literature reviews that have been done and those are important. Our consensus, by the way, was different from - a consensus conference in most other fields, the consensus conference members are the only ones who get to decide about the labels. That is, they don't go back into the field or go back into the discipline and get feedback. In most consensus conferences in medicine, for example, it's the attendees who come to a consensus and they basically say, "Here is our consensus."

John Bellone 53:32
Adopt it. [laughs]

Tom Guilmette 53:34
Right, exactly. So what we did really was different. We were much more transparent than what typically happens in a consensus conference in other disciplines. I'm so happy that we did because I think there's going to be greater buy in among clinicians. Clinicians will feel like - you know, psychologists are sensitive about these sorts of things, whereas other folks, other disciplines might not be. So I'm really glad that we went down this road because I think getting feedback from many different constituents within our organization, within our field, strengthens our ability to make these recommendations.

John Bellone 54:10
I like your process a lot. And then you mentioned a little bit about how to create a consensus statement. If Ryan and I think of a topic, which I'm sure we will. [laughs]

Ryan Van Patten 54:20
There are many.

John Bellone 54:20
We have plenty. So it sounds like you approach AACN about the possibility of forming a board and picking chair members?

Tom Guilmette 54:29
Right. Yes.

John Bellone 54:29
Yeah, okay.

Tom Guilmette 54:30
Right, exactly. So we went to them with a model, which was essentially based on the 2009 consensus conferences on malingering, and that was the model that we used. I can't tell you what their thought processes were except to presume that they have to see this as being an important enough issue within the field to get organizational support and sanction. Once we had written the consensus statement, which all 22 of us had to agree with - so this was, I mean, Jerry Sweet and I did the bulk of the writing, but then really everyone looked at [it]. First, the co-chairs of each of the three groups read through it, gave their suggestions. And then it was sent to everyone else. Then we had about three or four outside readers who also looked at this. This also then went to the Publications Committee of

AACN. They had to give their seal of approval, and then [it] had to go to the board of directors to get their final approval. So this was a multi-step process. So you obviously need significant buy-in on the part of the organization to be able to acknowledge that this is an important enough issue to go down this route and to commit resources to this, which basically means a boxed lunch and the conference space, at least in our case, the day before the AACN annual meeting.

John Bellone 55:53
Yeah, maybe it'll be a little while till we do that given your harrowing experience. But maybe a courageous listener will take up the helm on a topic. [laughs]

Tom Guilmette 56:03
You know, as labor intensive as this was, and at times frustrating, it was also and has been one of the most professionally rewarding experiences that I've ever had.

John Bellone 56:12
That's nice to hear. Yeah.

Tom Guilmette 56:13
I mean, you know, there were some major players in our consensus conference group, and everyone left their egos at the door. Everyone really embraced the fact that this is something that we need to do for the field. So it was very rewarding to do this.

John Bellone 56:30
Yeah.

Ryan Van Patten 56:30
It's great to hear.

John Bellone 56:31
All right, we've got a couple of bonus questions before we end. So the first one is: if you could improve one thing about the field of neuropsychology, what would that be?

Tom Guilmette 56:44

I think we still need to do continuing and ongoing work to enhance and update our normative data. I think it's important that we do more of this across measures so that we have a unified normative data set for different kinds of measures. I think we need to be really looking at getting better and better data for folks in the upper age ranges, now that people are living longer, to really know what is normal for a 90 year old or even a 95 year old. I think this really is the fundamental nature of what we do. Our ability to determine where a test score falls in a distribution of folks of similar age, education, ethnic background, cultural issues - our ability to make inferences about a score, and to derive diagnoses from scores, pattern of scores, as well as background history and all those sort of thing rests on the fact that we have adequate normative data. It's not sexy or glamorous, but I think it is critical to the nature of the kind of decisions that we make.

John Bellone 57:58

We're always hanging our interpretations on what the expected level of performance for that patient is. And that's what the normative data does. It gives you an expectation when you don't have baseline test scores for that patient.

Tom Guilmette 58:12

Right, for sure.

Ryan Van Patten 58:12

It's a precursor to our conversation today. If a score is low average or below average, z-scores of -1 or 1.5, that depends greatly on the raw score plus the normative comparison.

Tom Guilmette 58:24

Yes, right. For sure.

Ryan Van Patten 58:26

The second bonus question is what is one bit of advice you wish someone told you when you were training, or that someone did tell you that really made a difference? We're looking for an actionable step that trainees can take that they might not have thought of that can improve their training and performance.

Tom Guilmette 58:41

So we've already touched on this before, but I think one of the most important things that I was taught and made to really consider, and this incorporates some of the clinical judgment issues in our field, which is the risk of overpathologizing. Of interpreting normal variation as evidence of impairment by being too concerned with range of scores rather than pattern of scores. I know this is something that's near and dear to [your] heart, as well. It sounds pretty simple and basic, but, boy, that was something that, in some parts of my training, was really drilled into me. Even now, many years later, when I look at a profile and I'm thinking, "How far does this score or this pattern fall below expectations?" I am always thinking still about our inclination to overpathologize and our inclination to interpret normal variability as abnormal. That is something that I think we can't stress enough with trainees.

John Bellone 59:44

Yeah, we're always balancing Type I and Type II errors, right?

Tom Guilmette 59:49

Right, exactly.

John Bellone 59:50

Saying someone has Alzheimer's disease when they don't or saying they don't have it when they do, these are different errors that we're trying to avoid. There's a fine line to walk between those two. I agree that, in general, maybe we lean on one side too much and we should be cognizant of that.

Ryan Van Patten 1:00:09

We're making too many Type I errors?

John Bellone 1:00:10

Potentially. Yeah.

Tom Guilmette 1:00:12

I think trainees are prone to that especially because they - and maybe I'm going with our field and I'm just projecting my own history. [laughs]

Ryan Van Patten 1:00:22

[laughs]

Tom Guilmette 1:00:22
But if you're a trainee, you want to show your supervisor how much you know. So that may mean that, "I'm going to come up with a *diagnosis*. I'm going to show how smart I am. So I can make a *diagnosis*. Of abnormality, of impairment." I think we at times might sort of "value" some of that. But we might "value" that in terms of the supervisor/supervisee relationship that, "I can identify a disease. I can identify an impairment. I can identify a deficit," rather than calling it normal or saying that it's intermediate. I think trainees may have a tendency to want to impress by showing supervisors that they are capable of diagnosing neurologic dysfunction. And I think that might be to the error of considering the factors that we've just been describing.

John Bellone 1:01:25
Right. And patients sometimes find it more satisfying if you give them a diagnosis. I know sometimes people, surprisingly, are a little discouraged when I say, “The data looks great. I don't think there's anything going on right now.” But it's not validating their subjective symptoms.

Tom Guilmette 1:01:44
Precisely. Yes. Yeah, really good point.

Ryan Van Patten 1:01:46
You can feel anti-climatic, to do all this work - clinical interview, records review, so many tests, and then say, "You have a normal cognitive profile. Everything is fine."

Tom Guilmette 1:01:56
[laughs]

John Bellone 1:01:56
But that's the best outcome that we can expect.

Tom Guilmette 1:01:58
It is! Right. This is really easy. [laughs]

Ryan Van Patten 1:02:01
Right.

Tom Guilmette 1:02:02
I can summarize your functioning in two words, "All fine. All good". [laughs]

Ryan Van Patten 1:02:08
[laughs]

John Bellone 1:02:08
[laughs] Well, Tom, thanks so much. This has been a really great conversation.

Ryan Van Patten 1:02:14
Yeah. Thanks for all your work in this area.

John Bellone 1:02:16
True. Yeah.

Tom Guilmette 1:02:17

It has been my pleasure. Thank you so much for asking me. And also, thank you for the great work that you're doing on your podcast. It's fabulous. So keep up the good work.

Ryan Van Patten 1:02:23
Thanks. I appreciate it.

Tom Guilmette 1:02:24
You bet. Take care. Bye bye.

Transition Music 1:02:26

John Bellone 1:02:30

So that does it for our conversation with Tom. Before we end, we have just one point of clarification to make. During the discussion, we mentioned that neuropsychologists might use different test score labeling systems depending on the setting in which they practice and the populations that they serve. We actually were thinking about that a little more [and] we think that this test score labeling system can be used in all or virtually all settings. And that, really, it's the clinical interpretation that's going to differ depending on the particular setting. So we hope that that clears up any potential confusion. As always, thanks so much for listening. Join us next time as we continue to navigate the brain and behavior.

Exit Music 1:03:06

John Bellone 1:03:30

The Navigating Neuropsychology podcast and all the linked content is intended for general educational purposes only, and does not constitute the practice of psychology or any other professional healthcare advice and services.

Ryan Van Patten 1:03:41

No professional relationship is formed between us, John Bellone and Ryan Van Patten, and the listeners of this podcast. The information provided in Navigating Neuropsychology in the materials linked to the podcasts are used at listeners' own risk. Users should always seek appropriate medical and psychological care from the appropriate licensed healthcare provider.

End of Audio 1:03:59