Transcript | Validation Video 3.26.2026

SWYC Validation Training Video
3.26.2026
Transcript

People often describe some screening questionnaires as “validated”, and others as not validated. So what does it mean for a screener to be “validated”? That’s the question for this talk.

In my last talk, I discussed how we can think about screening like radar,
but we can also think about screening as a conversation starter. As I’ll describe, how we think about screening has implications for what kinds of research evidence are important—and thus, what it means to be validated.
often, we think of a validated screening questionnaire as one for which there is research evidence supporting its accuracy and reliability, as established through what are called "psychometric" evaluations. This is similar to asking whether radar can reliably and accurately detect airplanes. Make sense? Sure it does, as far as it goes. But saying a questionnaire is validated begs the question, validated for what? What is the screener intended to do?

For example, we might ask whether a screener is effective in various ways. A well-known model developed for diagnostic imaging that suggests a screener must first be

1. Feasible to implement, and
2. It must accurately detect the problems it is supposed to. So far, so good—sounds a lot like radar; but this model goes further by claiming that screening results must
3. Have an impact on clinical judgment. Unless someone's thinking is influenced by the results, the screener will have had no effect. In turn, the screener should display an
4. Impact on use of interventions. And finally, the screener should ultimately have an
5. Impact on children’s health. So to repeat, being “validated” only goes so far. To be effective, a screener must actually influence people. So let’s have a closer look at how typical research evidence on screening might influence clinical judgement and shared decision making (perhaps as a conversation starter), as well as how it applies to accuracy.

To do so, I’m going to ask you to imagine a screening questionnaire that yields scores between 0 and 9. Note that this is a simple model—a thought experiment, if you will. While it doesn’t correspond exactly to any real-world screener, I think you will find the resemblances to be informative. I’ll point out some key concepts along the way.

As with any screening questionnaire, Parents answer questions, clinicians add up the scores, and those scores fall over a wide range
We call this range of scores a distribution.
What can we say based on this evidence? For starters, we can say something about how common or rare a score is, ideally using plain language.

For example, if a child receives a score of 5, we can say that the child scored in the top 27.9%. Better yet, we might say that

Transcript | Validation Video 3.26.2026

1. out of 100 families, the parent reported more symptoms than about 72. Research supports plain language, and plain language facilitates conversation.
2. So here's our first key concept: when possible, use plain language to describe research evidence

If a child receives a score of 6, we can say that
1. out of 100 families, the parent reported more symptoms than about 91.

If a child receives a score of 7, we can say that

1. out of 100 families, the parent reported more symptoms than about 98. Already, this says a lot. If you agree that the questions are valid indicators of symptoms, knowing how extreme the scores are—and being able to explain that to parents—is clinically meaningful. Some assessment tools, such as the Child Behavior Checklist, define their clinical range based on data like these.
2. So I would say that key concept #2 is that normative data are useful--that is, its good to know how one person scored compared to other people. And I'll add that you don't need just one set of norms. For example, you could use national norms and also collect normative data from your clinic. That way, you could say something like, "out of 100 families nationwide, you reported more symptoms than 98. And compared to families in this clinic, you reported more than 85." Each normative comparison offers unique information.

But we can say more.

It is often useful to think about kids as falling into two categories. Here in blue is the first category—

1. those children who do not have developmental-behavioral problems, who I'll refer to as typically developing.

And here in red is the second category—those children who have developmental-behavioral problems. What can this tell us?

If the screener is supported by a good diagnostic accuracy study, we will know how many kids with a given score—in this case a score of 5—

1. Have a developmental-behavioral problem and
2. How many do not. Based on this evidence, we might say that:
3. "among children who scored 5 in a recent study, 27% were found to have developmental-behavioral problems"

1. "among children who scored 6,
2. 62% were found to have developmental-behavioral problems"

And "among children who scored 7,

1. 88% were found to have developmental-behavioral problems". This kind of information—i.e., probabilities associated with specific scores—is clinically useful. In my opinion,
2. That's key concept #3. You can get this kind of evidence in different ways. If it comes from a high quality diagnostic accuracy study, then you can be confident that the

Transcript | Validation Video 3.26.2026

screening score did not influence the diagnosis. If you get data from your own clinic about which kids end up being diagnosed, then you will want to be transparent about this, for example by saying that “among children with this score in our clinic, we end up diagnosing 88%.” Different information, but still clinically useful. As I'll discuss in a later talk, clinic-level data can be particularly useful for translated questionnaires, which often lack a robust evidence base. To understand evidence from a typical diagnostic accuracy study, consider the kids with developmental-behavioral problems.

So far in this example, they were folded into the overall distribution.

But we can pull them out

as their own independent distribution.

And we can look at their scores alongside the distribution for typically developing children. We can now see that,

1. typically developing children have an average score of 4,
2. whereas children with developmental behavioral problems have an average score of 5.5. On average, children with problems score higher—if they didn’t, the screener wouldn’t have any accuracy at all. But as you can see, there’s a lot of overlap

You can see the overlap even better if we flip the red distribution below the horizontal axis. Now we’re ready to consider the last element that makes this an example of real screening instrument—

1. the cut score, which is also known as a threshold, which determines which scores
2. indicate a positive screen
3. Versus a negative screen

Among kids with developmental-behavioral problems, this distinguishes

1. true positives – that is, kids with problems who we want to screen positive -- from
2. false negatives – that is, kids with problems who falsely screened negative. Likewise, among typically developing children, the threshold distinguishes between
3. True negatives – that is, kids without problems who should have screened negative – from
4. False positives – that is, kids who screened positive even though they had no problems

Here, I just shade the correct results—true positives and true negatives. Notice that a lot of the scores aren’t shaded in—there are plenty of incorrect screening results here. Let’s ask a few questions to see how accurate this screener is

1. Consider the children with problems who fall below the horizontal axis. What percentage are correctly classified as having a problem? That is, what proportion of scores fall in the “positive” range to the right of the threshold? Look at the figure and think about this for a moment. [click]
2. If you said “two thirds” or “three quarters”, you are very close. The actual number is 78%.

Transcript | Validation Video 3.26.2026

3. This is what we mean by “sensitivity”. Sensitivity is the percentage of children with problems who are correctly classified by the screener.
4. Now consider the typically developing children. What proportion are correctly classified as not having a problem? That is, what proportion fall to the left of the threshold in the negative range? Look at the figure and think about this for a moment.
5. Again, If you said “two thirds” or “three quarters”, you are very close. The actual number is 78%.
6. This is what we mean by “specificity”. That is, specificity is the percentage of children with NO problems who are correctly classified by the screener. This figure depicts scores on a screener that displays 78% sensitivity and 78% specificity. Most people would say this is a good screener. But is it good enough to make clinical decisions? Let’s take a closer look.

To think about this screening instrument from the perspective of a clinician, we might ask a different question. When a clinician cares for a child with a positive screen, the clinician would ideally like to know the chance that this child has an actual problem. [click]

1. So consider the children who score positive. These are the scores that fall to the right of the threshold in the positive of range, including those above and below the horizontal axis. Which positive scores are correct? That is, which scores co-occur with an actual problem? These are the scores below the axis. Among all children with positive scores, what percentage fall below the axis? Look at the figure and think about this for a moment [click]
2. If you said about half, then you are very close. In fact, less than 50% of children who score positive have actual problems. [click]
3. This is what we mean by positive predictive value, or PPV. So, if we implemented a policy that all children who score positive on this screener should be referred, then we would be wrong more than ½ the time, even though we are using a good screener to detect a highly prevalent condition.

But the situation is actually more complicated. We know that children with higher positive scores have higher risk—actually risk that is much higher than the PPV,

whereas others have moderate risk

And still other children—namely, those who score just above the threshold—are at much lower risk than the PPV would suggest.

The issue is that PPV averages risk across all who score positive. Yet we know that

The problem is that PPV averages risk across all who score positive. Yet we know that risks associated with individual scores vary widely. PPV may be useful for understanding groups of children with positive scores, but in the end, clinicians and parents tend to care about one child at a time. And each of those children has only one score at a time, and one particular prognosis.

Now remember, this is only a model. We should proceed with caution. That said, it is useful in at least two ways.

First, this model demonstrates that

Transcript | Validation Video 3.26.2026

1. tradeoffs are inherent in any clinical threshold—not just for developmental behavioral screeners – but for any medical test or any decision for that matter. As we raise the threshold,
2. Fewer children screen positive. This means that fewer children with problems screen positive, which reduces sensitivity but increases specificity and PPV. We see different tradeoffs when we
3. Lower the thresholds. More kids screen positive, which increases sensitivity but lower specificity and PPV. I'll discuss thresholds more in a later talk, but for now, let's consider a second point.

In short, this model tells us something about validated screening instruments. As I mentioned, sensitivity and specificity

1. Are 78% in this example. So this is precisely the kind of screener that is typically described as "validated." And yet,
2. 22% of children are misclassified. And if we believe the claim that Pediatricians should rely on validated screeners like this one when making referral decisions, not their clinical judgment, then we need to accept that
3. More than 50% of referrals will be false positives, with an even higher risk for children who score positive right at the threshold. Are validated screening instruments really sufficient by themselves to drive clinical decisions in this way? I think not, and that's
4. Key concept #4.

Now you may be thinking, ok fine, not all validated screening questionnaires can justify referrals, but the better ones can. Point taken. Maybe some screeners are highly accurate, like this:

And I'd say sure—that's possible, and if you can find me a screening questionnaire as accurate as that, I'd love to see it. But what I see more often is screeners
like the one I've been describing,

1. But without mention of predictive value
2. Without mention of errors at all
3. Emphasizing sensitivity and specificity, sometimes replacing it with some kind of
4. odds ratio between positive and negative screens (which is 11.4; often I see claims for screeners with odds ratios of 4 and under)
5. And certainly not including a visual of the distributions. As a result, I think that recommendations to use so called validated screening questionnaires in a foolishly consistent way—often based on limited information--is just another modern just so story. We can do better. For example, we can be clear about the limitations of even the most valid screening questionnaires.

And we can think more deeply about how we can use screening questionnaires to meaningfully inform clinical decision-making and conversations with families—ideally using plain language.

And this includes how we refer to our screening questionnaires. For example,
1. what's up with the use of the past tense here? Once the research done, is a screening questionnaire valid for all time and for all purposes, with no further room for debate? I think not.

Transcript | Validation Video 3.26.2026

The best screeners are supported by lots of high quality research of many kinds. But this kind of evidence should represent the beginning of a discussion about its best uses, not the end

I hope you this has been helpful. If you have more questions, check out our web site. And check back for updates because we hope to add new information in the future. Thank you for listening