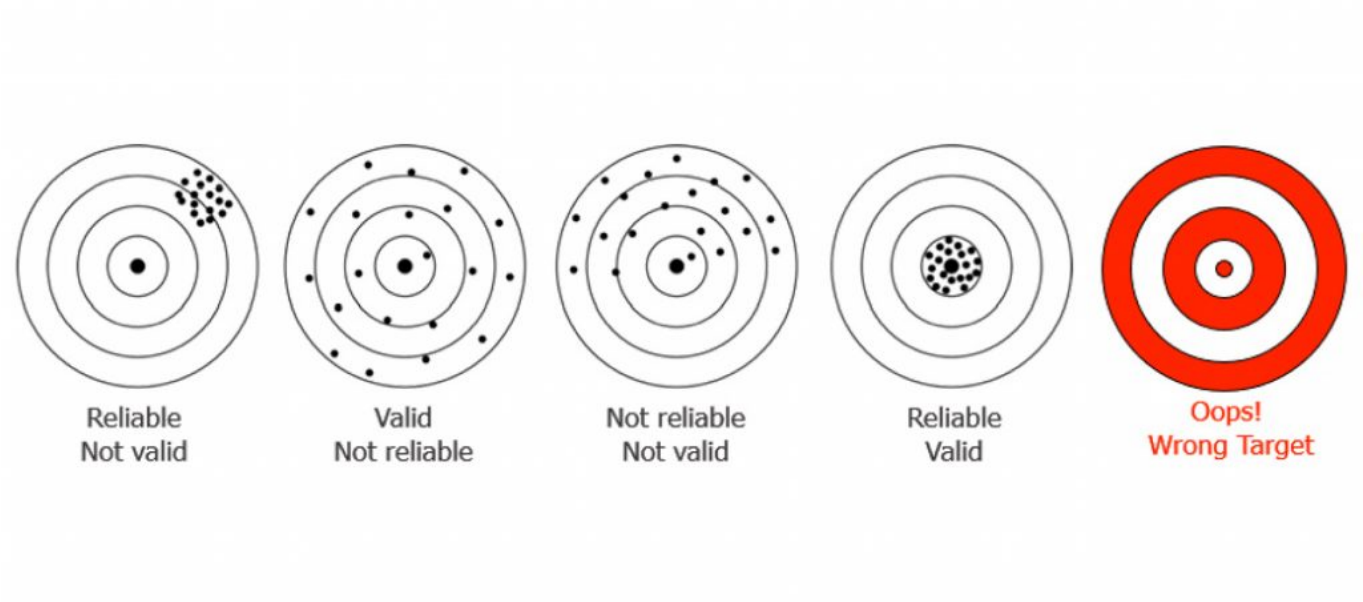
How to stay valid – in a measured way



"An important scientific innovation rarely makes its way by gradually winning over and converting its opponents, … What does happen is that its opponents gradually die out and that the growing generation is familiarized with the idea from the beginning."

Max Planck

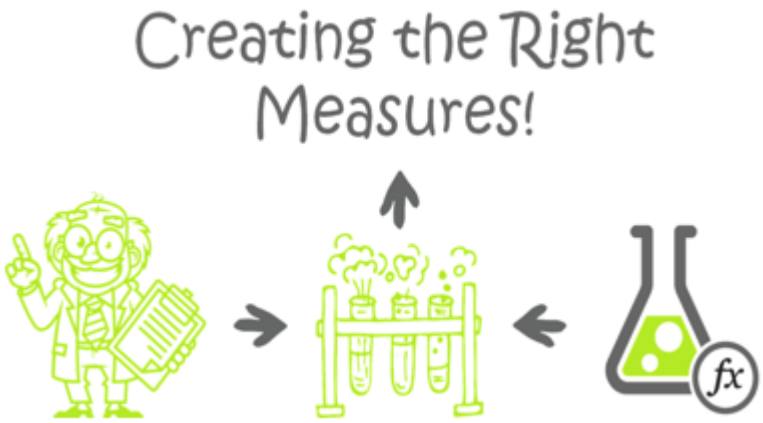
Recently I met with a client to discuss the potential use of behavioral assessments in their company. Much of the conversation was, as I anticipated, geared around scope, application and, more particularly, the broader issue of applicability (validity). As our discussion progressed, I started to get the feeling that we were somehow missing each other. As we prodded and poked the issue of assessments and how they are developed and what validation means I started realising that there was something fundamentally disparate in our viewpoints. I would cautiously attempt to address concerns while he would simply revert to an initial assumption discussed two or three issues earlier. Not to clarify, but re-insert a strongly held opinion. After one or two iterations, I got the message. Also, something clicked in my head. While we both, most likely, went through the same "age-appropriate" training on statistical validity in Psych 101, I realised that I was speaking about validity through a very different lens namely, Rasch measurement. The latest scientific innovation in the way we create behavioral measures. He in turn was still seeing things through a sixty-year-old lens where, a) we calibrate our test items by establishing how many people in a standard sample either succeed or endorses an item; b) item difficulty is defined by the proportion of endorsed or correct responses in a particular sample (p-value); c) the quality of an item is estimated from the correlation between the item response and the test score (point-biserial correlation) and finally; the person's ability is defined by the percentile rank in the respective sample. In essence, an approach totally dependent on the appropriateness of the sample used to standardise the test or assessment.

So, this puts me in a predicament. Here's a gatekeeper that has not as yet caught up with the fact that the way we traditionally validated scales no longer stands up to scrutiny. That we have since determined that a different approach is not only possible, but in fact the only way in which we can scientifically

develop human measures. And, that we now have an approach where we need make no assumptions about the people we used to develop the measure.

Regardless, we still end up religiously defending the assessments we've been weaned on despite them not even having the very fundamental measurement requirements of a zero calibration or a regular unit of measure!

However, the reality remains that many have still not figured out that the assessments we have so bitterly been defending and comparing everything else to does not even have the very fundamental requirements of a zero calibration or a regular unit of measure!



Yes, it's the way everyone did things for the last 60 years. But, we were wrong. We were counting and not measuring and in addition, looking for validation in all the wrong places. As

Benjamin Wright pointed out, *"The truth is that the so-called measurements we now make in educational testing are no damn good!"*

Consider this. From my client's perspective, and admittedly how I used to see things as well, when we administer a set of items to a candidate, for example Numerical Reasoning, we typically establish her Percentile Rank by comparing her to the group of people we standardised the test on. The fundamental problem with this, however, is the following; how do we understand our current measurement of this individual *beyond* the confines of the items we used to measure her and the group of individuals we originally used to standardise the test? Considering, the moment we change the group, we change the measure! Similarly, replace the items with other Numerical Reasoning items and, once again, you have a new measure!! Each collection of items will end up measuring an ability of its own while every person's score will be dependant for its meaning on a *specific group* of test takers.

So, while it's clear that the technicalities and thinking around developing and "validating" measures have fundamentally changed, I got the distinct feeling talking to my client that the breakdown was primarily around understanding that there has to be an acceptance that the traditional view of "validity" will look very different when we use Rasch measurement.

This article is really a shout out to my colleagues practicing in the people sciences to consider how we

have been measuring people and, more importantly, to resist not stepping back from our flat-earth perspective in order to consider other possibilities.

That being said, albeit in my own mind, I'd like to close the loop in terms of my discussion with my client by explaining how we need to reframe the idea of Validity given that the probabilistic Rasch measurement approach is the only way to develop scientific people measures. I'll do this in a way that contextualises how Rasch analysis of item response data explains, what we have for decades referenced in the Standards for Educational and Psychological testing, as **Validity**.

Counting is not measuring

Just before we look into the difficulty facing the way we traditionally approach validity, I'd like to briefly share with you what I mean when I refer to the concept of objectivity in measurement. This is important for us to understand why we need to reframe the way we view the concept of Validity of measurement instruments when we use the probabilistic Rasch measurement approach. You'll see where this fits in when I discuss Validity.

The effectiveness of how we measure human beings is completely dependent on our ability to transform our observations into measurement. When we apply physical measurement, as we do in science and our daily lives, we explicitly and implicitly use calibrated measurement systems. When we measure someone's temperature or measure the length of something, we don't give much thought to the "identity" of the instrument being used other than the fact that it is a *"member in good standing of the class of instruments appropriate for the job"*.

Figure 1

Sarah Haras

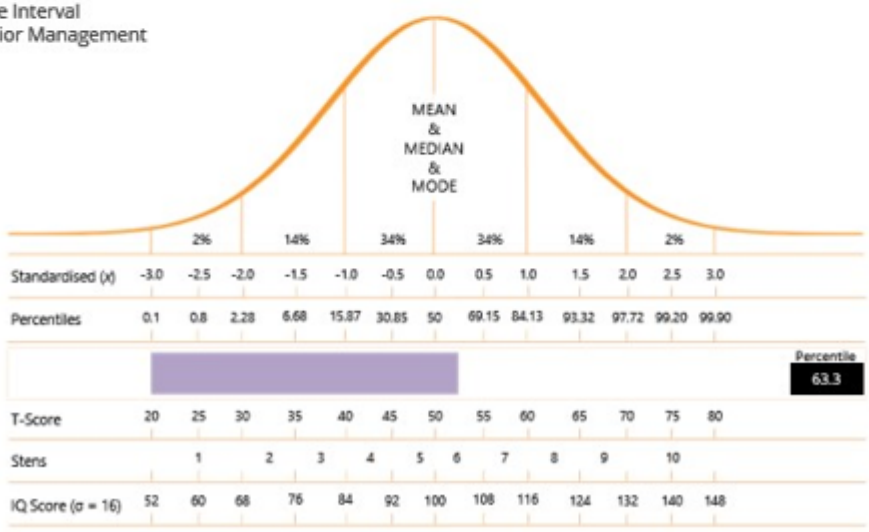
TAB®

Your TAB® Profile

NUMERICAL ESTIMATION	
NORM GROUP[2]	
Average Raw Score:	25
Percentile Rank:	45.5%
Accuracy Percentage Score:	97.2%
Standard Deviation:	3.12

INDIVIDUAL DATA	
Actual Raw Score:	26
Attempted Score:	28
Total Number of Questions:	30
Percentile Rank:	63.3%
Accuracy Percentage Score:	92.9%
± Standard Error of Measurement[1]:	0.70

[1]95% Confidence Interval
[2]Executive / Senior Management



Why is it then that when we determine that an individual, for example, has scored at the 63rd Percentile on a Numerical Estimation ability test, we immediately want to know *in what group* and *on which test*before we try to make any sense of it? How is it then that, if she tells us that she is 1.65 meters tall, *we don't immediately ask to see the tape measure she used to make that measurement?* Intuitively we accept that tape measures come in different colours, weights, and even lengths – concomitant criteria. However, we all assume they share a scale that is fundamentally independent of those associated characteristics and that they all will consistently give a measurement of 1.65 meters objective meaning. We reflexively also accept that any other person of the same height will more than likely measure 1.65 meters on any other tape measure. So, keep this in mind. Given the way we traditionally develop tests or measures, she will more than likely end up with a different Critical Reasoning ability Percentile Rank in every group she is compared to **however**, she will always be 1.65 meters tall in each of them! Like my client, this is where we all missed the boat. We perpetuated a people measurement technology that, by virtue of the way we constructed them, acted more like an elastic band than a standardised measure with common sense replicability – the way we measure things every day. So, why should I constantly be expected to "validate" assessment instruments against some other, often unrelated data set?

This is the measurement property we refer to as “objectivity”. It’s this property that underlies my fundamental premise for using Rasch methodology to create measures and redefine what we understand as validity.

As mentioned in my opening statement, for over sixty years now we’ve been calibrating test items simply by observing how many of the participants in the standard sample succeeded on a particular item. We then conclude that the particular item difficulty is the proportion of correct responses in the sample (ye old faithful p-value) and that the quality of the item is the correlation between the item and test scores (point-biserial correlation). Then, we make the inference that the person’s ability on the test is relative to their percentile rank in relation to the predetermined sample. This unfortunately relies blindly on the appropriateness of the standardised sample used? Equivalent to the height changing every time a tape measure is used to measure someone else of the exact same height?

How many still do it?

The Standards for Educational and Psychological Testing (“Standards”) is crystal clear that, Validity is “the most important consideration in test evaluation”. And that, in principle, is exactly the way it should be. However, the question now is more about what it means when it refers to *validity*? This is why the distinction between the Rasch measurement approach I now use and the traditional way I used to develop instruments, has to be highlighted. It is seminal to the assumptions we make around validity. As a principle, validity speaks to the way we take the results we get from administering an assessment or test and then how we ascribe meaning to the scores we get from the process. The following is pivotal. The Standards emphasise that it is the **inferences (interpretations)** that **are validated** and **not the test** and that only its use in certain contexts, for example for clinical applications, merits the designation of validity. For example, the Depression Anxiety Stress assessment is “valid” for clinical use. Obviously, no more than a simple “classification”?

The Standards I mention earlier reference three specific “types” of validity. Validity that’s Content, Criterion, and Construct related. Off the bat this should raise a flag given Validity, as defined by the Standards, is thought to be a “singular” thing framing an assessment’s legitimacy? We, however, have the Standards proposing these three facets as a combination of inputs into this single issue. The problem here is twofold. How do we reconcile these three proposed aspects of validity given each of them differ fundamentally in what they mean as well as the way in which they are applied? And, even if we do lump these three fundamentally disparate elements into a concept “validity”, how do we link the term to the analysis of the data? In the words of Wright and Stone, *“There are substantial and puzzling questions as to what is referred to, how it can be implemented in practice and what the results of implementation mean.”*

This is the dilemma facing our traditional approach to Validity. We cannot simply make the assumption that those three criteria are the primary and uniquely qualified inputs into what validity means. There will always be something else that needs to be considered. In fact, thinking any differently is almost

equivalent to the ridiculous Council of Nicaea's assumption that they had the divine right to decide what was to be regarded as the authoritative collection of the books of the Bible? When we decide on validity, in this context, how do we decide which of the criteria are indispensable, which are optional or, possibly a nice to have? We simply can't. Why? Because, as the likes of Cronbach, Campbell and Fiske suggested, *"Attempts to base validity on external criteria have raised more problems than they have solved."*

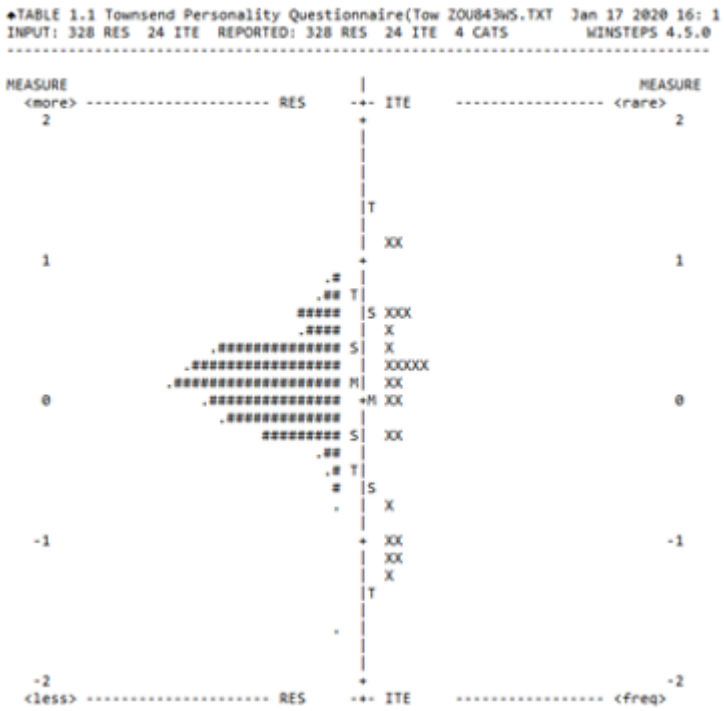
Is this what a flat earth is not?

How do we then avoid this dilemma? We have to shift our attention away from the Standards idea of validating the inferences made from the test scores, to the test data on hand – the *actual* responses the individual gives to the range of items. Only then do we ask, *"What is there in these data that could answer validity questions?"*

When we shift our focus to actual responses to the test or assessment items, it becomes clear that only two types of validity emerge. The first stems directly from the analysis of *the actual responses to the items*. It involves the way the items and individuals (their responses) are ordered and spaced along a clearly defined, single variable. This, in itself, assumes a well formulated and clearly defined understanding of the "thing" being measured. For example, Extraversion. Not a hodgepodge of unrelated "things" thrown together as we historically used to do. Only then to apply factor analysis, after the fact, to find out what congeals from the random pool of items. This is the first fundamental expectation of a scientific measure – linearity. In Rasch speak we refer to this as **Order Validity.**

How orderly are we?

Figure 2

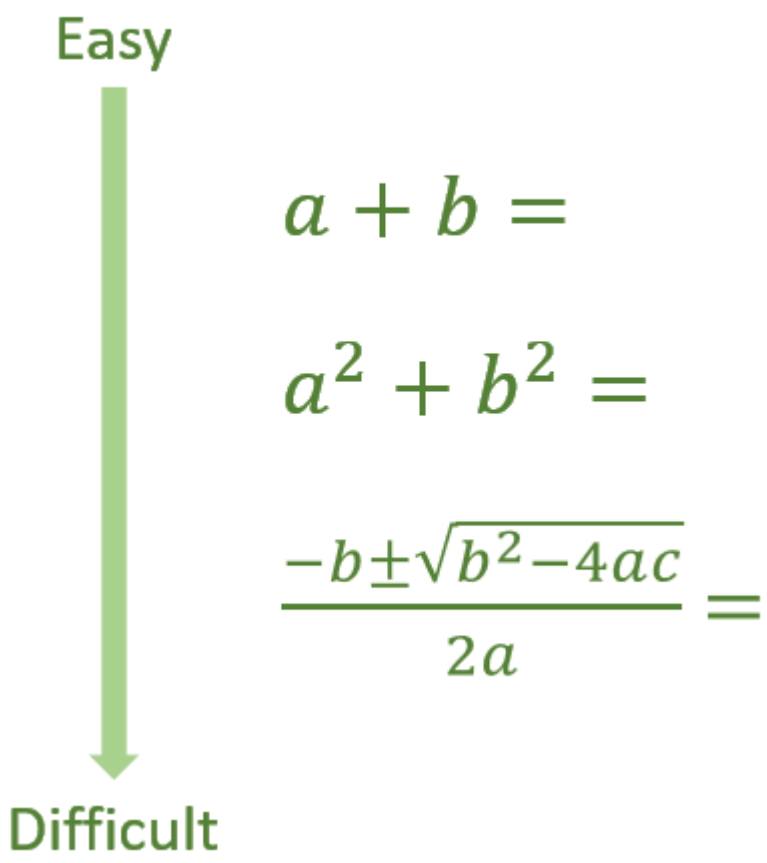


Now, to settle the possible paranoia generated by me dissing our bible, the Standards for Educational and Psychological Testing, let me say this. The Standards will always be our go-to for guidance on the criteria for the evaluation of tests, testing practices, and the effects of test use. The fact that it still frames the concept of Validity the way it does is, in my estimation, just a throwback to the last 60 years of how we “validated” instruments. Let me hazard a guess that soon, hopefully in the foreseeable future, Validity will be framed the way I and others are proposing here. As Order and Fit Validity based on the test data itself and not some arbitrary external criterion as I highlighted earlier.

Nevertheless, for those that are rightfully protecting the quality of the assessments and the inferences we make from them, this is where you’ll recognise the links you are anxiously looking for. When I refer to Order Validity, you should know that this operationalises what we historically referred to as Content and Construct Validity.

How we ensure Order Validity (≈ Content and Construct) with Rasch measurement, is by observing the way the person responds to an item of a clearly defined variable. For example, the Algebra Variable in Figure 3, by factoring in the item’s actual content and its relationship to the other items in terms of how much more difficult or easy it is to them. It is this relationship that we believe “either verifies, improves on, or contradicts” our intended contribution of the item to the underlying variable we intend to construct. By doing this, in our example, we establish the difficulty order of the items constituting our fictitious variable (Algebra Variable) and in so doing we define the variables meaning and subsequently, its so-called “construct” and “content” validity!

Figure 3



So, in future when I say my measures have Order Validity it should be code to my classically trained friends for Construct and Content Validity – wink-wink, nudge-nudge.

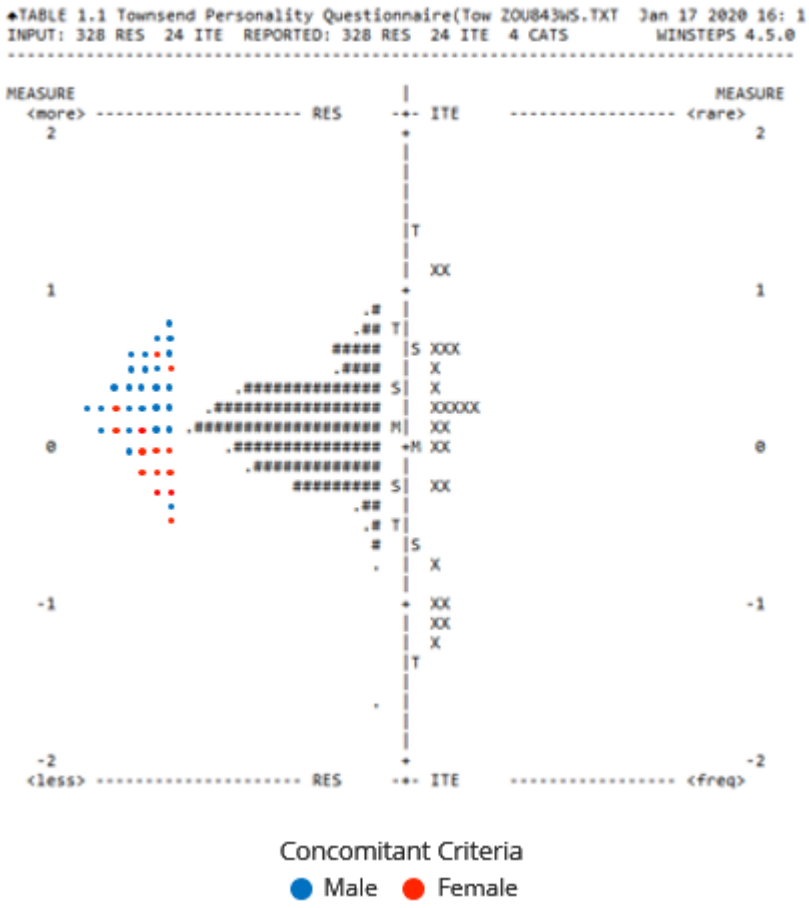
To summarise, "the difficulty order of the items defines the variable's meaning and hence its Content and Construct validity" and consequently, the way we order the responses of the individuals to the ordered items show the impact of them being measured on that specific variable which in turn, determines the variable's effectiveness.

And no, I have not forgotten our other classical friend, Criterion Validity. Where does this fit into this new way of doing things? Before I explain its approximation in Rasch measurement, let me just reiterate what I mentioned earlier in the article. We cannot simply make the assumption that the three-criterion specified for Validity by the Standards are the be-all and end-all. So, as I attempted to do earlier, I'll try to explain where that fits in to the current Rasch measurement thinking.

In the classical sense Criterion Validity as framed by the Standards, "presupposes the existence of an eternal criterion sufficiently well established to serve as the base against which the test can be compared." For those of you with a penchant for the more dramatic, this is where we usually deploy the correlation coefficient as an index to compare our existing measure to another measure that we assume measures the "same thing". For example, comparing Assessment A's Sociability dimension to Assessment B's Sociability dimension and then assessing the degree of consistency between the two. Seems simple enough. However, it's the very simplicity of how we identify the criterion that is problematic. Problematic

in that we have no idea whether the criterion we're correlating to is in fact valid. Or, does the comparison between these two test forms in fact add anything substantial to the assessment's validity?

Figure 4



I suggest that the classical idea of Criterion Validity is better served as a concept by interpolating these "concomitant criteria" along the variable map our Rasch model generates (Figure 4). In this way, we can have criteria like gender, age, education, job level, etc., plotted along the same underlying variable we're measuring together with the item calibrations as well as the person measures!

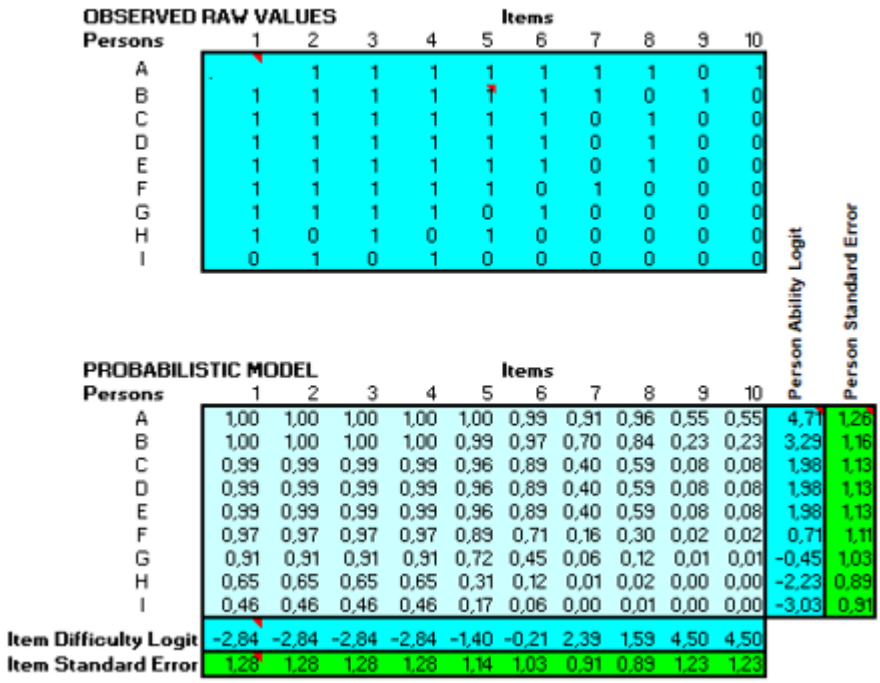
By doing this, we can start building a definitive picture of any specific criteria in relation to the item calibrations and person measures as opposed to simple correlations that more often than not just indicate the presence of some general relationship without exposing any potential underlying complexity. Since, as we all know, while having one foot in boiling water and the other in ice cold water is a perfect correlation for feet in water, it is most definitely not an optimal situation to be in.

Like with breaking up, fitting in is hard to do

Hopefully I've contextualised why, going forward, I no longer reference the three monoliths as laid out in the Standards for Educational and Psychological Testing namely, Construct, Content, and Criterion validity but rather refer to the concepts of Order and Fit validity. So, asking me to justify the use of my instruments designed using Rasch methodology, by using sixty-year-old technology, is the equivalent of having to prove that a GPS gets me to my destination the same way a road map used to. I just have to

point you to the strained relationships during holiday trips to appreciate that, *it does not*.

Figure 5



OBSERVED RAW VALUES

Persons	Items 1	2	3	4	5	6	7	8	9	10
A	.	1	1	1	1	1	1	1	0	1
B	1	1	1	1	1	1	1	0	1	0
C	1	1	1	1	1	1	0	1	0	0
D	1	1	1	1	1	1	0	1	0	0
E	1	1	1	1	1	1	0	1	0	0
F	1	1	1	1	1	0	1	0	0	0
G	1	1	1	1	0	1	0	0	0	0
H	1	0	1	0	1	0	0	0	0	0
I	0	1	0	1	0	0	0	0	0	0

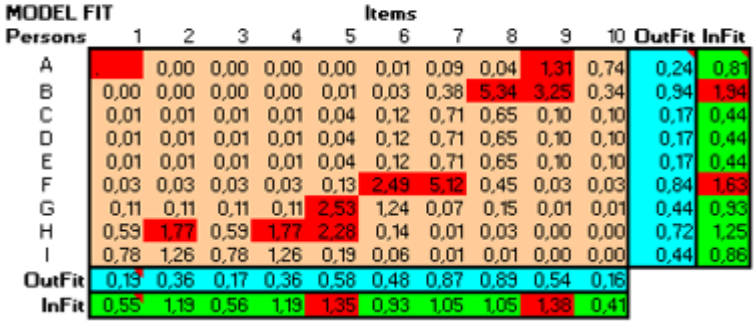
PROBABILISTIC MODEL

Persons	Items 1	2	3	4	5	6	7	8	9	10	Person Ability Logit	Person Standard Error
A	1,00	1,00	1,00	1,00	1,00	0,99	0,91	0,96	0,55	0,55	4,71	1,26
B	1,00	1,00	1,00	1,00	0,99	0,97	0,70	0,84	0,23	0,23	3,29	1,16
C	0,99	0,99	0,99	0,99	0,96	0,89	0,40	0,59	0,08	0,08	1,98	1,13
D	0,99	0,99	0,99	0,99	0,96	0,89	0,40	0,59	0,08	0,08	1,98	1,13
E	0,99	0,99	0,99	0,99	0,96	0,89	0,40	0,59	0,08	0,08	1,98	1,13
F	0,97	0,97	0,97	0,97	0,89	0,71	0,16	0,30	0,02	0,02	0,71	1,11
G	0,91	0,91	0,91	0,91	0,72	0,45	0,06	0,12	0,01	0,01	-0,45	1,03
H	0,65	0,65	0,65	0,65	0,31	0,12	0,01	0,02	0,00	0,00	-2,23	0,89
I	0,46	0,46	0,46	0,46	0,17	0,06	0,00	0,01	0,00	0,00	-3,03	0,91
Item Difficulty Logit	-2,84	-2,84	-2,84	-2,84	-1,40	-0,21	2,39	1,59	4,50	4,50		
Item Standard Error	1,28	1,28	1,28	1,28	1,14	1,03	0,91	0,89	1,23	1,23		

Before I conclude let me just briefly draw your attention to a second type of validity this new way of developing people measures considers. This type of validity has to do with the consistency of both the response patterns of the individuals to the set of items. Essentially, it speaks to how well the response pattern from the individual aligns with the items along the clearly defined variable being measured. Also referred to by Wright and Stone as the “response performance validity for persons and items”. Exactly as explained and satisfied by L. L. Thurstone and yet, now for obvious reasons highlighted above, not mentioned in the Standards. Not because the Standards are incorrect in any way but, because the criterion underlying the approach to measurement was never based on developing a measure but rather applying intricate statistics to nothing more than counts. So, understandably, Thurstone specifying the importance of Fit Validity as a necessity when developing measures simply could not be entertained because the underlying premise simply could not accommodate the idea of “Fit Validity”. Consequently, the Standards absence of fit statistics is symptomatic not of their lack of a desire to get things right, but simply because there is no underlying model for what we expect to measure against! A “lack of awareness of what we are trying to do”.

My client is probably thinking, “But what about point-biserial coefficients? We’ve been using that as item fit statistics for decades!” The problem again is that, like all of us previously, none of us had a clue as to what the statistical model for point-biserials is? We had, and still have, no idea what size coefficients we need to be looking out for or, in fact, to act on.

Figure 6



MODEL FIT	Items											
Persons	1	2	3	4	5	6	7	8	9	10	OutFit	InFit
A	.	0,00	0,00	0,00	0,00	0,01	0,09	0,04	1,31	0,74	0,24	0,81
B	0,00	0,00	0,00	0,00	0,01	0,03	0,38	5,34	3,25	0,34	0,94	1,94
C	0,01	0,01	0,01	0,01	0,04	0,12	0,71	0,65	0,10	0,10	0,17	0,44
D	0,01	0,01	0,01	0,01	0,04	0,12	0,71	0,65	0,10	0,10	0,17	0,44
E	0,01	0,01	0,01	0,01	0,04	0,12	0,71	0,65	0,10	0,10	0,17	0,44
F	0,03	0,03	0,03	0,03	0,13	2,49	5,12	0,45	0,03	0,03	0,84	1,63
G	0,11	0,11	0,11	0,11	2,53	1,24	0,07	0,15	0,01	0,01	0,44	0,93
H	0,59	1,77	0,59	1,77	2,28	0,14	0,01	0,03	0,00	0,00	0,72	1,25
I	0,78	1,26	0,78	1,26	0,19	0,06	0,01	0,01	0,00	0,00	0,44	0,86
OutFit	0,19	0,36	0,17	0,36	0,58	0,48	0,87	0,89	0,54	0,16		
InFit	0,55	1,19	0,56	1,19	1,35	0,93	1,05	1,05	1,38	0,41		

This is why the Rasch model, and nothing else, allows us to use the raw scores (percent correct) attained from the responses to the items as necessary and "sufficient statistics for estimating person abilities and item p-values" for estimating item difficulties (Figure 5). This enables us to use the actual data from the response set to develop a probabilistic model (Figure 5) which we in turn use to establish whether our data fits the model (Figure 6).

This, mathematically, "specifies what kind of relationship must be approximated between the **observed data** and the **estimated measures** in order for **valid** calibrations and measures to result".

A take-away thought

Unfortunately, and not because people don't care, many folks are carrying this unnecessary "validation" yoke and, like good custodians, defending it for all the right reasons. After all, when we set out to "quantify" people we cannot have everyone randomly prodding and poking away at them without any checks and balances?

So, what to do?

Well, I hope in this short article I've at least made a case for everyone that is vested in accurate and reliable people measures, to start relooking the way we've always done things and, with just a bit of healthy scepticism, possibly build on (or replace) what our colleagues in years past so creatively gave us with the limited resources and world view they had at the time.

For more information or assistance, please contact Gary Townsend at Skillworx Africa (Pty) Ltd, mobile 0824915392, gary@skillworxafrica.com. Or feel free to browse our website on www.skillworxafrica.com.