education review
a journal of book reviews

Volume 10 Number 5 **April 26, 2007**

A Little Less than Valid: An Essay Review

Noel Wilson
Lenswood, South Australia

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. (2002). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

$49.95 ISBN 0-935302-25-5

Citation: Wilson, Noel. (2007, April 26). A little less than valid: An essay review. *Education Review, 10*(5). Retrieved [date] from http://edrev.asu.edu/essays/v10n5index.html

As a test maker I worked for the Australian Council for Educational Research for six years. As a result I had always regarded this book in its previous incarnations as a sort of bible, a reference of last resort. So not until I wrote my Ph D thesis on Educational Standards and the Problem of Error did I subject the 1985 version of *Standards* to a more critical analysis (Wilson, 1997). As that analysis was not overly complimentary, I thought it only fair to look at the 2002 version with similar critical gaze. As before, I focus on validity. Why? Because, as the good book says,

> Validity is, therefore, the most fundamental consideration in developing tests (p. 9).

I concur. If the test event is not valid, if indeed the test is invalid, then all else is vain and illusory.

Validity

So what is validity?

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed use of tests. . . the proposed interpretation refers to the construct or concepts the test is intended to measure (p. 9).

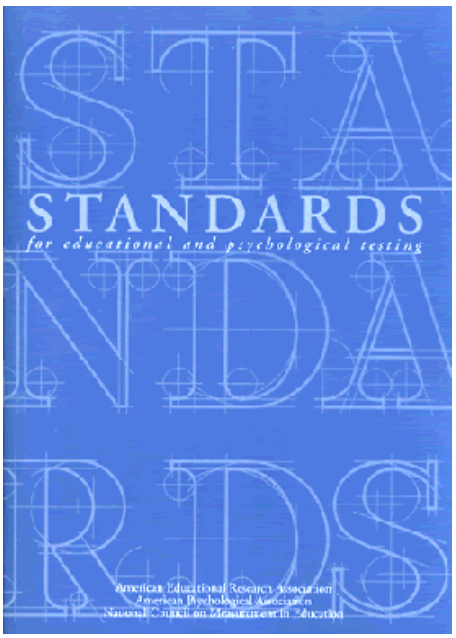
Further to this,

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses (p. 17).

That is, the validation argument presents a cogent case for the defence. Such an argument becomes a professional guarantee that the whole event of testing is valid. Indeed, this is spelt out.

> Ultimately, the validity of the intended interpretation of test scores relies on all of the available evidence relevant to the technical quality of the whole system. This includes evidence of careful test construction; adequate reliability, appropriate test administration and sorting; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees . . . .
> (p. 17)

The encouraging part of this is that validity is now seen as a function of the total test event and of the interpretations that surround it. As we shall see, *Standards* (2002) is very selective about which elements of the test event it chooses to focus on, and is equally selective about which interpretations it is willing to attend to. Even so, it has accepted the principle that it is the total test event that is involved here, not just the test, or just the scores, or just the hypothetical construct the scores are deemed to measure. So when I claim later that the score, or the classification, cannot be attached to the test taker, because it belongs to the total testing event, I would expect supporting grunts from the writers of the 2002 *Standards*.



Now let's look at the dark side of the truncated definition of validity here described. When I went to

school, validity was about whether an assessment measured what it purported to measure. And what were estimates of the errors. Invalidity was assumed as an empirical fact. The problem was to estimate its extent.

But times change. According to *Standards* (2002), to show that a test is valid, what you are required to do is to show that you have examined in detail some aspects of validity, and can produce evidence that in these aspects some glimmer of validity exists. I say glimmer because there are no real standards in this book of *Standards* about what, for example, even an acceptable level of reliability might be in particular cases. And of all validity aspects, reliability is the one most beloved by psychometricians, and most lauded by test constructors. So what level of reliability constitutes an adequate standard? That's a matter for professional judgment.

What does all this mean in practice? Surely that any test can be validated, and thus deemed valid, because not only are those involved in the validation process not encouraged to address all sources of test event invalidity, but are positively encouraged not to.

> Professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretations and use. As in all scientific endeavours, the quality of the evidence is primary. A few lines of solid evidence regarding a particular proposition are better than numerous lines of evidence of questionable quality. (p. 11)

The advocacy described earlier is again evident here. The reference to scientific work is fatuous. Genuine scientific endeavours examine all of the evidence. It is legal endeavours that favour advocacy bias.

The real problem with this definition goes deeper. The test event does not include the validation process. The validation as described is the supporting argument for using the test data in particular ways. It is an adjunct to the test event, and separate from it. So whether the validation is or isn't done, the test event remains unchanged. Logically then the empirical test validity, to what extent the test event does what it claims it does, is independent of the validity argument. It is clear that the test validation is a public relations spin for the test event. The empirical validity or invalidity is undisturbed by the validation process. And we are back to square one in our invalidity discourse. What is the extent of the invalidity error in the total test event?

And that's a problem, for another direct result of *Standards 2002*'s truncated definition of validity is its eradication of the word 'invalidity' from the discourse. In its chapter on validity, the word 'validity' appears 60 times (plus or minus 3), and the word 'validation' 23 times (plus or minus 2). The word 'invalidity' appears on three occasions, in connection with consequences of testing (p. 16, p. 24), and here only to limit collateral damage, as it were, by indicating that unless the negative consequences of testing can be shown to have a direct link to 'a source of validity such as construct underrepresentation or construct irrelevant components,' then it "falls outside the technical purview of

validity." (p. 18) I think that what this says is that the only admissible possible sources of invalidity according to *Standards 2002* are construct underrepresentation and construct irrelevance.

So much for the 13 sources of invalidity described by Wilson (1997), and the fifty plus sources referred to by Messick (1989).

On further investigation, I'm not sure that even these aspects of invalidity aren't discounted by the glossary entry on Construct Validity.

> In the current Standards, all test scores are viewed as measures of some construct, so the phrase (construct validity) is redundant with validity. The validity argument establishes the construct validity of a test. (p. 174)

So all tests measure constructs by definition. The validity of the test is established by the validity argument, which is a set of assertions. And the name of the construct is the name of the test.

The map is the territory. Asserting something makes it true. General Semantics is herewith abolished.

In regard to consequences, surely by its own definition of validity this contraction of invalidity regarding consequences is unacceptable. For "interpretations of test scores" surely include the practical consequences of those interpretations, the point of contact of the interpretations with the test taker. And such interpretations, at least in educational testing, inevitability have both social and psychological effects. The focus tends to be on the positive effects, on those who "succeed," however that is construed. But the social consequence for many is one of exclusion, of being denied assess to certain further selection processes or occupations or studies or whatever. And the psychological consequences for many are indeed harrowing, especially if the implied "failure" label associated with educational testing becomes both repetitive and acceptable, and thus finally incorporated into the construction of the self.

Let's go back and look more closely at this narrowing of the invalidity components to construct underrepresentation and irrelevance. Irrelevant to what? Primarily the construct that the test measures, which in practice is the essential definition of the construct, as determined by the test items. And the name of the test, of course, is the name of the construct. The circularity of the definitions is mind boggling, and the testing discourse spirals inwards. In fact, the major source of invalidity error is not to be found in the translation of a particular test constructor's notion of a construct into test items or other performances. Rather it is to be found in the translation of what is supposedly being measured into the construct. It stems from the lack, indeed impossibility, of clear definition in the educational or psychological world as to what in clear empirical terms is the thing to be measured, and then in the gap between this description of the required test taker behaviour, and the behaviour required when the test taker scores on the test. And from the point of view of the test taker, the greatest test irrelevance resides in the very

form of the test itself, independent of content. Many test takers simply do not accept that this test performance is related much to what she knows, or can do. Such considerations are essentially excised by the narrowed and controlled definitions in the variables involved, such as construct, test format, and measurement error.

The text of *Standards 2000* is constantly slippery on issues such as these. Here is an example.

> Nearly all tests leave out elements that some potential users believe should be measured and include some elements that some potential users consider inappropriate. Validation involves careful attention to possible distortion in meaning issuing from inadequate representation of the construct. . . . [T]he process of validation may lead to revisions of the test, the conceptual framework of the test, or both. (p. 10)

Let's unpack this. The first statement is a statement of fact. Let's accept it as true. To me it implies that the construct has a variety of interpretations, one of which has been chosen by the test maker, probably because it is amenable to test making. The second sentence talks of "distortion of meaning" and "inadequate representation" of the construct. The slipperiness resides in the unstated intervening epistemological assumption that there is one true description of the construct which differences help us to resolve, rather than accepting the more obvious evidence that the "construct" is multi-layered with soft edges, thus necessarily ill-defined and ill-definable. As with test scores, different meanings of a construct are not deviations from the true meaning, but different and adequate and alternate legitimate descriptions. There is a faith among professional examiners that a technical fix can resolve real differences in acceptable meaning. It can't.

Reliability

At this point I want to spend a little time on the reliability issue, for if there is a particular psychometric villain in this cover up of the real extent of invalidity within the world of professional testing, if there is a Judas lurking, then reliability, that most precious of indicators, on which so much attention is lavished and so much theoretical rigor and empirical error data amassed, reliability must surely be it. How can this be, when reliability admits error, is indeed awash with it?

First, let's be clear about one thing. Reliability is a sub-set of validity. Putting it another way, of the thirteen sources of invalidity listed by Wilson (1997), one deals with instrument error, which is essentially what reliability error is. The chapter on Reliability claims to have much more centrality than that.

> To say that a score includes a component of error implies that there is a hypothetical error-free value that characterises an examinee at the time of testing (p. 25)

I chose to commence with this sentence because it so brilliantly sums up the simple plethora of assumptions on which the whole structure of educational and psychological testing is premised, and on which its current practice relies. It asserts the logical necessity and empirical reality of the "true score" or its counterpart as a measurable attribute of the individual test taker. A truism hardly worth commenting on, were you thinking?

Yet the statement is completely fallacious. All measurements contain error. It is the very nature of measurement. And it does not imply a "true" value. It simply implies that no measuring instrument is perfect and no measuring procedure is perfect, so measurements of the same thing will differ. It is error that differentiates a measurement from a definition or a standard, which is without error. The 'hypothetical error-free value' is not a deduction from anything. It is an assumption of a statistical theory. And this hypothetical score does not characterise an examinee, it characterises a complex testing event which includes the examinee. Or more accurately, includes a group of examinees. To attribute any score to the examinee is an attachment error. And, of course, it is not possible in an individual case to make any comment about how far from the estimated score the hypothetical true score might be. For reliability data is group data, and it is a logical type error to apply such data to individuals in the group.

The notion of reliability goes deeper in its deceptive practice. Reliability involves statistical theories whose symbols have no necessary correspondence in the social world. The symbols cannot be assumed to have real world counterparts. But when such mathematical symbols are given labels such as 'true score', and 'trait and ability parameter', then these counterparts are asserted to be present in the human psychological and social world. And when the test is given a title such as 'problem solving' , or 'literacy', or 'practical mechanics', or 'American history', then this true score or ability becomes defined further by the name of the test. And when this verbal label is then attached to the test taker, this purely statistical parameter becomes reborn as an attribute or deficiency of an individual test taker's persona, pinning him or her in the appropriate place, the correct rank order, on the competitive specimen butterfly board. During this magic transformation, the measure has broken free from its invalidity associations, yet still manages to retain its lineage of mathematical genesis, now claiming scientific status, and professional (because of the hypothetical validity argument) impeccability. Impeccable because to the professionals and users the measurement error, which is always acceptable whether it is measured or not, has become the total error in the testing event, and to the general public and the students because there is no error, and the estimate has transmogrified into the true score, securely attached to the victim's psyche.

And where does this scam begin? By defining errors of measurement in terms of reliability, instead of in terms of validity. This is asserted in large print in the very title of Chapter 2, Reliability and Errors of Measurement, and then measurement error is defined.

> The hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true or universe score for the procedure is called measurement error. (p. 25)

And true to form, it becomes what it is called.

A huge sleight of hand has been accomplished. The error in the individual test taker's score has been colonised by psychometricians. Suddenly the measurement error has been narrowed to the world of a particular set of test items or questions or performances in a particular test situation, and all other aspects of invalidity effectively relegated to the wastepaper basket.

And now the plot thickens. There is worse to come.

Oh what a tangled web.

There is another problem with reliability. *Standards 2002* concludes that in regard to current movements in educational testing,

> each step towards greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or construct underrepresentation in an assessment program. (p. 26)

Let me unpack this grudging admission. For the acceptance of the inverse relationship of reliability to validity is indeed grudging. Note the use of "may," and the usual limitation of invalidity effect to the relatively narrow issues of construct underrepresentation and irrelevance.

I have detailed elsewhere (Wilson, 1997) precisely how the mechanisms used to increase reliability do result in an increase in invalidity in regard to many of its aspects. Here are some examples.

Temporal errors are maximised by obtaining only one single score at a single time. Humans learn, and forget, and make mistakes. So test behaviours, and hence test scores, will change over time. These differences in estimate describe the temporal invalidity. Largely ignored in the chapter on validity, they are discussed in some detail in the chapter on reliability in *Standards 2002*. While their importance is indicated, there appears to be no obligation to determine its effect on either group norms, nor on individual test takers.

Contextual invalidity is increased when assessment is limited to a single pencil and paper situation and to the very artificial environment in which "reliable" testing occurs. Contextual error includes all those differences in performance and its assessment that occur for other methods and contexts for obtaining relevant data.

Construction invalidity is likewise indicated by the different assessments obtained that are not constrained by the limitations of content, form, process and media contained in usual testing and examination procedures. Again the capacity to generalise, and thus the validity, is diminished by the psychometric strictures required for high reliability.

What I had not realised before writing this essay was how much of the more ontological invalidity aspects were directly attributable to the assumptions and implicit demands of the psychometric models in use, of the fudges necessary to maintain them, and the subsequent unsustainable claims about what the categorizations mean and what the testing event was able to accomplish. So the invalidity of the testing is due not so much to the inadequacies of the test as such, as to the claims generated by reliability theory about what the test can do. For remember, invalidity is a description of the extent to which the test cannot do what it purports to do. The greater those claims, the greater necessarily is the invalidity. So, in practice, what does reliability theory purport to do in terms of its own definitions, logic, and fudges?

You're asking what are the implicit and explicit claims made through reliability theory?

That's right.

Well, it claims there is a true score for each person who does the test.

A true score of what?

A true score of what the test measures.

And what does it measure?

A single construct or ability or trait.

How do we know that?

Because the test items all relate to that construct.

And?

And because they assert it.

But aren't most constructs multidimensional?

Almost certainly.

So aren't the items in this test put together in this now unitary construct in a very idiosyncratic way?

They probably are.

So what's special about this particular construction of the construct?

It's what the test measures?

You mean it's the true construct?

That's a pretty strong implication.

And what's the name of this true construct that the test measures and that has a true score that is also a trait or ability of the test taker?

That's obvious. It has the same name as the name of the construct that the test was designed to measure and that is the name of the test.

How do you know?

That's what the test makers and users claim.

It must be hard to prove that all those claims are valid.

That's true. Much easier to prove that they're not.

Cut-off scores and categorisations

The test taker has the most to gain and the most to lose by the interpretation of the testing event, and in particular of the interpretation of her particular test score. In this sense the test taker is the major stake holder in the test event. And to the stake holder it is the report, which contains the categorisation based on his performance in regard to the "construct" that is the permanent attachment. Not permanent because it describes some achievement or ability or trait, but because the number or symbol on the report describes the categorisation that, in high stakes testing, predicates his future.

Categories require a cut-off point that defines the standard of adequacy. If there is an accurate scale, any measure can be measured as over or under that rather arbitrary and prejudged cut-off. But psychometric tests are different. The scale does not exist till after the test event is completed. Despite the claims of some practitioners of item response theory.

Given its importance, the attention *Standards 2002* gives to this crucial element in the testing event is negligible. There is no mention of cut scores in the chapter on Validity. Yet surely the greatest determiner of invalidity has to be the miscategorisation of any particular examinee.

There are two paragraphs in the chapter on Reliability. The first indicates that

> Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 35)

Further to this,

> When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure using the same form or alternative forms of the instrument. (p. 35)

This is sturdy stuff. I look forward to seeing this information in the public arena in the near future.

The only other mention appears in the chapter on scales, norms and score comparability. The discussion on how precisely cut scores might best be determined is very general and trivialises both the importance and complexity of the issue. (pp. 53 & 54). As, indeed, do the standards that derive from them. However, some comments on the standards are informative.

> With achievement or proficiency tests such as those used in licensure, suitable criterion groups, (e.g., successful versus unsuccessful practitioners) are often unavailable. Nonetheless it is highly desirable, when appropriate and feasible, to investigate the relation between test scores and performance in relevant practical

> settings. Note that a carefully designed and implemented procedure based solely on judgments of content relevance and item difficulties may be preferable to an empirical study with an inadequate criterion measure or other deficiencies. Professional judgment is required to determine an appropriate standard setting approach (or combination of approaches) in any given setting. (p. 60)

So while it is most desirable to do some reality testing with criterion measures in 'relevant practical settings', we are warned, twice, that these may be 'unavailable' or inadequate. So it may be preferable to stay within the closed world of reliability checks, and not venture into the dangerous world where invalidity lurks. And I wonder to what extent criterion measures are deemed inadequate because they have low reliability. And possibly high validity?

It seems obvious to me that the test taker should be warned of this parlous state of the art around the matters of cut scores, categorisations, and credentialing.

As a starter therefore, I have drafted a short explanation that might be appended to any report of a categorisation of a test taker who engages in a testing event.

> WARNING: Takers of this test should be aware that, despite the high reliability of the test, we estimate that a minimum of 20 percent [modify as appropriate] of test takers have been miscategorised by the test because of measurement error and of the arbitrary and unstable nature of the cut scores. The upper limit is unknown. The test is, however, free of unfairness or bias, for these effects are random, and as many test takers have been over-categorised as under-categorised. Unfortunately we are unable to determine into which group this particular test result falls.

Collections of estimates

While educational testing and psychological testing may have different professional purposes and intentions, their social and political uses are very similar. Both are crucial elements in the categorisation of the people who take the test. Such categorisations are then used to select and exclude for a range of occupational courses and futures.

To the extent that these categorisations are accurate or valid at an individual level, these decisions may be both ethically acceptable to the decision makers, and rationally and emotionally acceptable to the test takers and their advocates. They accept the judgments of their society regarding their mental or emotional capabilities. But to the extent that such categorisations are invalid, they must be deemed unacceptable to all concerned.

Further, to the extent that this invalidity is hidden or denied, they are all involved in a culture of symbolic violence. This is violence related to the meaning of the categorisation event where, firstly, the real source of violation, the state or educational institution that

controls the meanings of the categorisations, are disguised, and the authority appears to come from another source, in this case from professional opinion backed by scientific research. If you do not believe this, then consider that no matter how high the status of an educator, his voice is unheard unless he belongs to the relevant institution. And finally a symbolically violent event is one in which what is manifestly unjust is asserted to be fair and just. In the case of testing, where massive errors and thus miscategorisations are suppressed, scores and categorisations are given with no hint of their large invalidity components. It is significant that in the chapter on Rights and responsibilities of test users, considerable attention is given to the responsibility of the test taker not to cheat. Fair enough. But where is the balancing responsibility of the test user not to cheat, not to pretend that a test event has accuracy vastly exceeding technical or social reality? Indeed where is the indication to the test taker of any inaccuracy at all, except possibly arithmetic additions?

We only break free from this long historical tradition of structural and symbolic violence when we acknowledge openly the huge invalidity or error components in any categorisation of individual people to all stake holders, including those most severely affected, the test takers.

To do this appropriately we must turn our backs firmly on theoretical statistical systems, which have been at the heart of the problem. Instead we must turn to genuine empirical data. What is required are independent estimates of the categorisation, or the data on which such categorisations are premised, from as many legitimate sources as possible. We aim for a collection of data that in most cases, for most people, will be classified as unreliable, in that there will be very significant differences in the estimates. Not because of the inability of the estimates to be "scientific" or "objective" enough to get even close to the "true score." But because the estimates, collectively, dispel the very notion of a "true score," and indicate the variety of legitimate estimates of the "constructs" or "displays of human response," which differentially emphasise aspects of the multidimensional aspects of both construct and performance to which the estimates are a response.

Again, it is important to dispel the notion that this collection of estimates represents a random variation whose mean is the "universe score." It doesn’t and it isn’t. Any attempt to fiddle statistically with the independent estimates is a move to destroy data to achieve an ideological purpose – that there is a single ‘true’ rank order of something measured by a test event, or a collection of test events, so it may be established that it is fair competition that produces winners and losers. These are not the words of a profession, but of a political ideology.

Here’s an example of a collection. Students present a portfolio of some examples of their school work to an external authority. The portfolio might contain four pieces of work which have already been assessed according to some scale or other, by one or more people. The four pieces represent different aspects of their learning, of their work. Each piece might also have been categorised according to some future predictions or

provisions. The collection might receive further independent assessments from the external authority, including estimates of an overall categorisation.

What happens at this point is crucial to the whole validity issue. What we have is a collection of work and assessment and categorisation estimates. This is the truest, most valid picture of the student we are able to obtain. It is, we might expect, disparate and subjective. Any apparent objective measures have, as ever, been produced initially and finally by subjective selection and interpretation and judgment. Any summation or averaging or other statistical manipulation of these data not only destroys their integrity, but perpetuates a myth about accuracy, and disempowers the student by robbing him or her of access to the information on which any final categorisation must be based.

For of course this information should all be made available to the student. To claim that the student may misinterpret these complex data is fatuous. The student may equally misinterpret simpler data such as the final categorisation judgment by accepting its implication of infallible accuracy.

Professional assessors, of course, might look at a collection of estimates such as this and shudder with horror. Because from their perspective this is a picture of hell. A picture which displays invalidity, error, in all its abhorrent expanse.

Yet critical reflection about the collection of estimates, all different, of measures of what the student knows and can do, reveals that these differences are not really an indication of invalidity of the test event at all. Why should they be invalid just because they are different?

For the whole notion of invalidity as error, as difference, is premised on the notion of validity as the true score, which has zero error, and on the construct being measured as the true and valid construct, so that any deviation (under or over representation) is error. And finally it is premised on the notion of this true score on a true construct being attached to the test taker as a true and stable measure of a trait or ability which magically has the same name as the construct.

Let's go back again to the old-fashioned definition of validity as the degree to which a testing event can do what it purports to do. Most of the invalidity in a testing event based on the assumptions of the last paragraph is not due to incompetent construction or marking of tests or categorisations of test scores. They are not due to technical problems, so cannot be fixed by technical fixes. They are due to the ontological problems embedded in the assumptions themselves. When these assumptions disappear, what the test purports to do changes, so the major sources of invalidity disappear.

The collection of estimates is the best picture we can get of an inherently incorrigible set of human performances designed to elucidate some learned patterns of behaviour, seen from different perspectives. As such it represents a true or valid picture, but a picture that in its essence is permeated with differences, with contradictions. The collection then

presents the test taker as a genuine human, not a static number or category on a scale, but a paradox in motion.

A short historical epilogue

In the beginning is the politics. The political task is to maintain good order through good ordering, through the categorisation of the person, through the professional selection for privilege and exclusion. Scientific theory and the competitive ethic will ensure both fairness and capitalist correctness in this endeavour.

Professional judges are inadequate for this task. They are unreliable and unfair. So was psychometrics born, and reliability became its good child. Invalidity was the bad child who was hidden in the closet while her look-alike, born again construct validity, was slotted neatly into the vacant space.

Science required measurement, and measurement required error. Reliability theory produced an error score based on random variations from a hypothetical mean. This mean was called the true score, and error called not instrument error, but measurement error. Thus was validity error neatly bypassed.

But what is this true score measuring? It has to be a unity of some sort. So let's call it a construct out there. The unity of what the test is designed to measure. We'll give it a name, the same name as we give to the test. And the same name that we give the construct that construct validity is all about.

Now watch carefully. We assert that this construct also exists as a permanent entity inside the psyche of the test taker, with the same name as the test name and the construct out there. And the quality of that inside construct is defined as its quantity as measured by the true score, or more precisely as its best estimate which is the number of correct items. This quietly though, for the implication that the estimate is the true score makes a better story.

So the inside circle is closed. Everything is defined in terms of something else inside the circle, and we are able to produce an apparently error free order which can be used to rank order and thus categorise test takers.

Only problem is, the whole point of the exercise is to define the cut-off points that will define the standards. Without these there can be no categorisations, and the whole political exercise (charade) has been futile.

How define the cut off points, the standards? Professional judgment seems to be the answer, as it always seems to be when the technical answers dry up. And so it is, and the outer political ring has also come full circle. But didn't we start with the professional judges being unreliable and unfair?

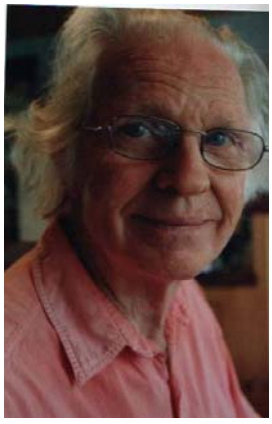
Only one question still remains to be answered about this terribly self-contained system of scientism, ontological fantasy and deceit. To what particular set of professional standards do the professional judgments ultimately defer? Education, Psychology, Psychometrics, Administration, Accountancy, Law, or Politics?

References

Messick, Samuel. (1989). Validity. In Robert L. Linn (Ed.). *Educational measurement (Third edition)*. New York: American Council on Education, Macmillan Publishing Company.

Wilson. N. (1997), Educational standards and the problem of error. Education Policy Analysis Archives, 6(10). Retrieved April 25, 2007, from http://epaa.asu.edu/epaa/vol6n10/.

About the Author



Noel Wilson is an educator, researcher and writer who lives in the Adelaide Hills in South Australia. He still writes stories which search in vain for publishers. He is delighted that his mind works better now at seventy five than it did at forty. Every now and then he has a little foray back into the educational field. He is a long odds optimist because he believes that sooner or later schools and educational systems will get better. And he'd be pleased to engage in dialogue about this essay.

He can be reached at noelwilson26@hotmail.com.