The Rasch Model: Empowering Educational Achievement in the Developing World

Trevor G. Bond
tgbond007@gmail.com
James Cook University, Australia

ABSTRACT: *This paper contrasts accountability focussed Rasch based assessment systems with those that are aimed at supporting student learning, by drawing at distinctions between assessment of learning and assessment for learning. It includes a modest proposal for implementing a Rasch based assessment system in developing countries. After highlighting problems with current educational testing regimes, it describes two exemplary assessment systems which provide good models for developing testing in developing countries. In particular, it canvasses specific recommendations which move away from testing of learning to testing for learning.*
"If I had to reduce all of educational psychology to just one principle, I would say this. The most important single factor affecting learning is what the learner already knows. Ascertain this and teach him accordingly" (Ausubel, 1968).

KEYWORDS: **Rasch model, Rasch based assessment systems, educational achievement, developing countries**

Received 30/05/2023 ➔ Revised manuscript received 30/07/2023 ➔ Published 30/12/2023.

1. Introduction

This paper is a personal attempt to draw inferences from my experiences as an educator and assessment consultant to advise on the development of testing and assessment programmes in developing countries. So, in part, I want to draw a distinction between those Rasch based testing programmes that hold schools and teachers and children accountable, and those that are aimed at supporting learning. Many educators might think that if you have a Rasch based educational testing system, then, *ipso facto*, you must have the best sort of educational testing system. But, my colleagues and I will report that often when we consult with executives on their school testing programs, we will be told: *"Oh, our programme is so wonderful, it's just so wonderful. We put the learners at the center of the testing, we put the teachers at the side - by doing this with testing, and doing that with testing."* And we walk away, shaking our heads, because a key feature of all that they're doing is trying to hold people accountable: students, teachers, parents and administrators, all held accountable.

So, we need to distinguish some differences between assessment of learning, which focuses on the past, and assessment for learning, which focuses on the future. In order to do that, we can reflect on some of the problems that we have already with educational testing systems; yes, even those wonderful, Rasch measurement-based systems, before looking at two much better systems from two different places in the world. This will emphasise why we should be moving away from testing of learning, towards testing for learning.

One key principle has persisted with me from the second year of my undergraduate teacher education programme: *"If I had to reduce all of educational psychology to just one principle, I would say this. The most important single factor affecting learning is what the learner already knows. Ascertain this and teach him accordingly."* (Ausubel, 1968). If you are a Piagetian, that's the rule; if you are a Brunerian, that's the rule; if you follow Skinner, that's the rule. If you follow Vygotsky, that's the rule. The most important thing affecting learning is what the learner knows now. Start there.

2. Some shortcomings of educational testing

There's been a very persistent and longstanding criticism of what was one of the most important testing regimes in the world - the

Graduate Record Exam - which allowed people to matriculate into universities in the United States. For half a century, it has been impossible to know whether the standard of that test was the same for years on end. When administrators were challenged with the idea that the GRE standards were declining over time, it was impossible for them to demonstrate that they were not; because the data were not collected in a form which allowed for making those over-time comparisons (Wilson, 1988).

Even a Rasch based national assessment program, such as the one we have implemented in Australia, NAPLAN, has a number of crucial weaknesses. The first is, there is a long time-gap between when the students are tested and when their teachers and others get the results. If you believe in a concept of education, you know that the child in 6-months' time is not the same child as was tested 6 months earlier. Secondly, it is impossible to track any particular child's progress over time, because NAPLAN is designed to test at the classroom or system level, not to test at the individual level; the standard errors of measurement at the individual level are so large that any changes in scores over time would, most likely, be hidden. And anyhow, the system was never designed to do that. The system, unfortunately, was designed to keep teachers accountable, to keep schools accountable, to keep state education systems accountable. And quite frankly, much better advice to the Australian government of the time might have been something like: *"You might think that's the national testing system you want, but we can tell you there is a better system. And that's a system that measures children's achievement over time, and gets results back to teachers quickly."*

As an unfortunate consequence, the NAPLAN test has become a high-stakes test in Australia's educational systems, just in the same way that the Rasch-based Territory Wide Assessments changed in Hong Kong. The Territory Wide Assessment in Hong Kong was introduced specifically, to be low-stakes diagnostic testing to inform teachers what was going on with the children in their classrooms. And almost immediately, and this is a reflection on the particular role of testing in Hong Kong, the TWA went from being a low-stakes test to a high-stakes test. Early in my time as professor in Hong Kong, I read of a little lower primary girl in the New Territories jumping out of the window of her parents' apartment block to commit suicide, because she had 'scored poorly' (something over 80%) in the last 'low-stakes' examination she took.

So, the key question here, then, becomes: Are there systems that can work better for students and teachers, and for learning? What can we see that is distinctive about them? And then how can we make that work in practice? For developing countries: if you're going to copy something, copy the best. Don't copy the rest.

In reminding you of the Ausubel (1968) claim *"ascertain what the learner already knows and teach him accordingly"*: assessments assess what the learner already knows, then teach that child based on what you just assessed. The focus shifts from the past to the future; from summative assessment to formative assessment; from assessment of learning to assessment for learning. My claim is that Ausubel was saying that 60 years ago, but he did not use those terms. He didn't say *"assessment of"* and *"assessment for"*. But 60 years ago, he was telling us what assessment should be about: it should be finding where the child is, and then saying, let's start teaching there, and move forward.

An analogy from health practice

Our present and future need is to put the focus on the teacher, on the learner, and the teaching-learning relationship. We need to know where the student is now. And we need to inform the teacher about that immediately. Take an analogy from health practices. You go to your GP and report that you have some health issues. The GP responds by giving you a blood pressure test, collecting a urine sample, and a cursory ENT check. In conclusion, the GP sends away for some other histology/pathology tests. And your GP gets the results of those scientifically calibrated tests back in 2- or 3-days' time. The GP's initial clinical diagnosis is confirmed/modified/contradicted, and the appropriate treatment/remediation programme is implemented.

Just imagine a medical system where, instead, the whole population is tested once a year on a standardized health test battery, then, some six months later, the nation's GPs' reports say: Here are the problems your patients have. What are all these obese people doing in your practice? You are not a very good doctor, you have got all these grossly overweight people; look at all these smokers in your practice, hypertensives, and so on. But that is exactly what we do with testing in many educational settings. The medical professions would never stand for that. Yet teachers have to deal with that all the time. Moreover, teachers do not have daily access to calibrated and standardised testing as the medical professionals do. Most of their testing is self-constructed, unvalidated, and never calibrated. Just imagine the furore if medical practitioners' blood pressure machines, thermometers and scales were removed from their consulting rooms. That's right, teachers do not have adequate access to the equivalent of such basic standardised instruments as these. And, most importantly, teachers need to assess whether the student progresses over time. Does the child grow? That's more important than does the child pass or fail?

3. Growth - the core issue in education

In order to do that, we need to have teachers, students and families on the side of assessment. But that is not how we implemented NAPLAN in Australia. NAPLAN was imposed on the education system by the federal government. It's a series of tests focused on basic skills that are administered annually to every single child in grades 3, 5, 7, and 9 right across the country. By way of contrast, in Vietnam, authorities implement purposeful sampling of the children, and then use those results to infer, to deduce what is happening nation-wide.

Imagine, as a parallel, a large commercial fishery system where, in order to determine whether the fishery was working well, they had to pick up every single fish, and weigh it, every year, then throw it back. That testing is traumatic and not conducive to growth; weighing does not make the fish grow faster. NAPLAN gives a 2-year snapshot of how children are performing in reading, spelling, punctuation, grammar, and numeracy. Not any particular child, but children, in general. So, NAPLAN has a system wide focus on change. Table 1, which I have chosen rather tendentiously, is from a recent national NAPLAN report, and shows NAPLAN achievement trends over a decade and a half in numeracy, reading, writing, spelling, grammar, for Years 3, 5, 7, and 9 across Australia. It summarizes whether achievements, on average, remain the same, (e.g., numeracy and reading for years 7 & 9), improve (e.g., spelling for years 3, 5 & 7), or vary. Neat, but how is that going to inform the teaching you intend to do with the child/ren to whom you teach reading in Grade 5 tomorrow. And, what if you teach one of the off-year grades? (See the NAPLAN website for the range of reporting formats.)

Table 1. NAPLAN Achievement Trends 2008 - 2022

	Year 3	**Year 5**	**Year 7**	**Year 9**
Numeracy	—	/	—	—
Reading	/	/	—	—
Writing	—	∪	∪	∪
Spelling	/	/	/	∩
Grammar	⟨	—	—	—

/ Average trending up 2008-2022; ⟨ Trend positive but flattening; — Average unchanged 2008-2022; ∪ Early downward trend reversing in recent years; ∩ Early upwardward trend reversing in recent years. (After the NAPLAN report for 2022)

4. Alternative Rasch-based assessment systems

Are there other systems? Where can we look? We can compare the imposition of NAPLAN in Australia with the way in which two other Rasch model-based assessment systems have been implemented elsewhere: e-asTTle, a teacher operated assessment system in New Zealand; and, MAP, from NWEA in Portland in Oregon, where schools buy into the assessment system provided by a not-for-profit assessment organization.

4.1. New Zealand: the e-asTTle system

So, how was the asTTle system implemented in New Zealand? A conservative government proposed national transition point assessments (for entry to intermediate/high school) and locked in funding in the budget for that. However, that government lost the next election without calling for tenders.

The subsequent, more liberal, government called for tenders and a team led by Professor John Hattie proposed an item bank for diagnostic test usage that won the hearts and minds of the Curriculum division of the Ministry. So successful was the roll-out for Years 4-8 in primary schools, that the secondary teachers' union asked for it in secondary schools, to create norms for years 9-12 across curriculum levels 5/6. The immediate result of this unique collaboration in New Zealand was a classroom-based assessment system, from which teachers could find out where their kids were at any time relevant to the outcomes of the national curriculum, by sitting at their desk-top computers with a supplied testing CD, developing tests, scoring tests, and interpreting the results.

When do they get the results? As soon as they put the data into the machine. Teachers could give the test on Friday, take it home and mark it on the weekend, enter the answers into the machine on Monday, and press the button. Individual student results were available straight away, along with all the national comparisons, district comparisons, like-school comparisons, and so on. Most teachers now use the online system that automatically scores the test and populates a report menu system after all students in the relevant group have been tested. School leaders can then look to see what was happening nationally in schools. The e-asTTle system covers years 5 to 12, but as long as your pupils are somewhere on the curriculum, you can use it if you want to find out where those children are now.

Schools find it useful for planning purposes; the important thing is the teachers and learners, and parents of those learners can find out where the child is now, and what can be done for that child, today. Teachers help the students (and parents) to understand what it means to progress, or *"what did I do?"* The routine school report default, *"I scored a B this year, a B next year, and a C after that."* carries no sense of progress, growth, or development. What does that mean? We want to know where the children start, to put an achievement ruler beside each child's development, and say where the child was last week, say where it is at the end of the year, and say where it is in June, next year.

What makes the e-asTTle system different? The teacher can use the testing system at any time. All the teacher has to do is sit down in front of the computer and click on, 'I want a test.' And the computer dialogue box replies, 'What sort of test?' The teacher interacts with the computer to move indicators across the screen to allocate grade level, subject, content area, knowledge components, and so on. Teachers can't choose or omit individual questions, but they can enter the specifications for the testing. The test is then developed according to the teacher's specifications by the algorithms that are built into the testing system. If the teacher decides that the generated test doesn't really match the requirements of that classroom, today, then the teacher can reject the test and start again by reconfiguring the test specifications. There are thousands and thousands of possible tests. Each time a teacher asks for a test, e-asTTle delivers a new test. Re-entering the same specifications in a week's time will result in a new test, different from the test a week ago.

One of the interesting by-products of this testing, was how teacher discussion changed when presented quickly with calibrated test results. Teachers started talking to each other about how it was some had good maths results, but others had poor. By contrast, that teacher's class achieved the highest language results. Just imagine the new staffroom chats: Oh listen, let's do a swap. You teach all the maths in both classes, and I'll teach all the language. Or, my class results have a big gap here with regard to the introduction to algebra; what do you do in your class to understand algebra, when I can't seem to do it with my class? Would you share your materials with me? Would you like to take

my class for remedial algebra sessions? Can you share your ideas? And I can then implement them in my classroom? And, rather than just having a score out of say, 40, teachers find that the output identifies students' weaknesses, and, also student strengths.

E-ASTTLe testing results are presented in accessible, visual formats, so teachers can communicate those results back to the children, their parents, and to school leaders. And the results are tied to a link with teaching resources, What Next, and specifies the curriculum demands and what learning experiences would be appropriate for those students. The leading psychometrician of the team that put these ideas into practice, Prof. John Hattie, was most noted for his three-parameter IRT modeling. When Hattie reported on the early implementation of the asTTle system (when the teachers were supplied with a CD for implementing asTTle in their classrooms) to a group of researchers at AERA, he, quite unexpectedly for me, pointed out that the whole system was based on the Rasch model. Moreover, he said, teachers could not be convinced that there's any more important information than the total number of items that the child gets right. In Rasch measurement, N is the sufficient statistic for ability and difficulty estimation. Now, that's been the key part of classroom teaching and learning. That's how it's always been done; Rasch measurement put into practice the commonsense ideas that teachers have of what testing should be like. And teachers would not listen to 2PL and 3PL modeling.

It's called e-asTTle now, of course, because it's all done electronically and online. Results show the distributions of scores with mean, standard deviation and the range. More importantly, it shows via an adapted Wright Map which learning objectives were easy vs hard, and right vs wrong, for groups and individuals. So, you can see how pupils grow across the programme from beginning to the end of school. Graphs reveal how the scores are linked to the curriculum levels, and that the curriculum actually moves up the same ability scale. So, the shift, then, is from assessment of learning, to assessment for learning; from summative to formative assessment. The purpose is to assess what the student already knows and what they might learn next. And this information is provided directly and immediately to the teacher. But what's more important than the total score is how that can be interpreted in a learning context. You need to have 22 scale points to make a measurable difference in asTTLe scores. So, if the children move more than 22 points - between two students, or two classes, or two performances over time, that's measurable growth. If the difference is less than that, no meaning can be attributed. (See the e-asTTLe website for the range of test specifications and reporting formats.)

4.2. Northwest Evaluation Association: MAP

The remarkable fact is that school systems voluntarily buy in and administer NWEA MAP assessments to pupils on top of the state and national testing that is legislated by the authorities. This is because what school administrators and their teachers see as being helpful to document where the children are now, and whether those children are likely to graduate appropriately at the end of their education, if they continue on their current achievement trajectories. Just imagine any school system, and no school system ever has enough money, then spending its money to do additional testing, just because that additional MAP testing is focused directly at the teachers, and the learners, and their learning.

The Northwest Evaluation Association started what has now become MAP. How did that eventuate? George Ingebo was in a curriculum assessment group in Portland, Oregon, in the USA. He had heard that Benjamin Wright in Chicago was making what seemed to be ridiculous claims about testing and measurement. The very dubious Ingebo went to see Ben Wright in Chicago and sat in on his workshops. When George Ingebo came back to the Northwest, he advised that the NWEA had to implement Rasch measurement to calibrate curriculum demand in order to find out whether the curriculum was aligned with growth. NWEA's focus moved very quickly from assessment of the curriculum to assessment of learners. (Ingebo, 1997) Then, more gradually, the focus went from assessment

of learners to assessment for learning. It's a not-for-profit organization, and it is a huge operation: nine and a half thousand schools and districts, now in 145 countries. MAP assessments help teachers identify student needs. They track mastery and measure academic growth over time. Tests are given to students at prescribed times, three times a year in order to help the educators plan a local curriculum that matches current student ability.

How is this different from the New Zealand approach? It's the same test for everybody, apart from the constraint that the tests are delivered in line with the local curriculum requirements. The difference is just local curriculum, not the idiosyncratic requirements of any particular teacher for that day's classroom assessment. Testing is centrally administered online with calibrated results available almost instantaneously, adopting informative methods of visualizing student educational progression. Then, the core MAP ethos is a focus on assessment for learning, and the idea is very simple: see their needs, close the gaps, help them grow; i.e., assessment for learning. Progress maps, calibrated on the RIT-scale, are provided for individuals, classes and other groups of interest. Projections of likely success on exit exams are plotted, based on the child's most recent assessment and growth to that date. And, what is a RIT scale? A Rasch logIT scale. (See the NWEA website for the range of MAP reporting formats.)

5. Educational testing in a developing country

Educational assessment systems in a developing country (Dang, 2022) are likely to include classroom and school-based assessments, national assessments, and then formal examinations which allow students to move on at important times in their transitions from one level of schooling to another.

Among the key benefits of using Rasch measurements for educational testing, both the items and the persons can be estimated on a single interval measurement scale. One of the most important things teachers could currently do to improve their assessments would be to look as closely at their items as they look at their students; unfortunately, even the teachers I know don't do that because their practice-based items are beyond scrutiny: only the children need to be modified or changed. A developing country could gradually implement standardised testing. In other words, construct tests to a standard across the school, across the district, across the country, and implement those standardised tests in a particular testing situation. This allows/requires item banking. It is encouraging and surprising to see that local university colleagues in Vietnam have already had some success in building, from scratch, an item banking system and implemented it for testing year 10 maths achievement here (Nguyen et al., 2022).

Why would it be important for a developing country to build its own item banking system? Because the price of buying into an educational testing system, a CAT - Computer Adaptive Testing system - is prohibitive; western testing companies are eager to contract out such systems, but control and expertise usually remain in the hands of the owners, not the users. Now, the locally built platform not only has item banking going, but it has computer adaptive testing on the way as well. These two attributes are central to the idea of measuring against standards, centrally important components for a developing country.

And, most importantly, we can work towards maintaining our assessment standards over time. We might matriculate more students to university in one year and fewer, the previous or next year, because of the number of students in the cohort who meet the standard, instead of accepting a fixed percentage.

So, here is a modest proposal for an imaginary developing country. Involve the teachers in developing an item bank for each aspect of their school-based assessments. Teachers already have these items to hand, locally available, all the time. Ask them to share with a centralised item bank. How do you make it worth their while? Give every teacher a code, so that the code for the item writer is attached to the item every time the item gets used; the teachers can reflect on/be acknowledged for, their own contributions. That way, the teacher becomes an active professional

partner in the testing process, not merely an accountability target of that testing process. Items should reflect the competencies that are required at every classroom level. For the teachers, item improvement would be an additional bonus, because, after receiving test feedback about the items, the contributing teachers/item developers would better understand what makes items (and distractors for MC tests) more or less difficult. Moreover, the focus on classroom competencies is interesting because the developing country in which I am interested has just implemented a competency-based educational system.

Next step, calibrate those competency-based items, and you need Rasch measurement to do that. All you have to do is to use those items, by administering them to classroom samples of pupils nearby, get the results, put the results into your item bank, give some more tests to other classes, but always include some common items between the tests. Continue until you can confirm the calibrations of the items on the assessment scale, and so on. Allow teachers and schools and districts to access printed tests for their school assessments. The school district might request, say, a grade 3 maths test, a grade 6 maths test and so on; the most straightforward place to start is with mathematics. This is because the development of mathematics items is much more straightforward - even across cultures - than the development of, say, language-based testing items. The vast majority of mathematics items are not going to involve issues of interpretation, as we often have with language. The school calls up the central system and requests a 30-item maths test for grade 3. The teacher receives the PDF file, prints it off, and then uses it in the appropriate classrooms. The teacher completes the assessment task in the usual way: √ X √ √ X √ X X √, and so on; and the data are shared back to the central system. After the teacher administers the test, scores the papers, Rasch analysis is used to provide feedback to the school for diagnosis and learning. And, gradually, then you can compare the strengths and weaknesses in a class, compare the students' results, the class result to the year level competencies and so on.

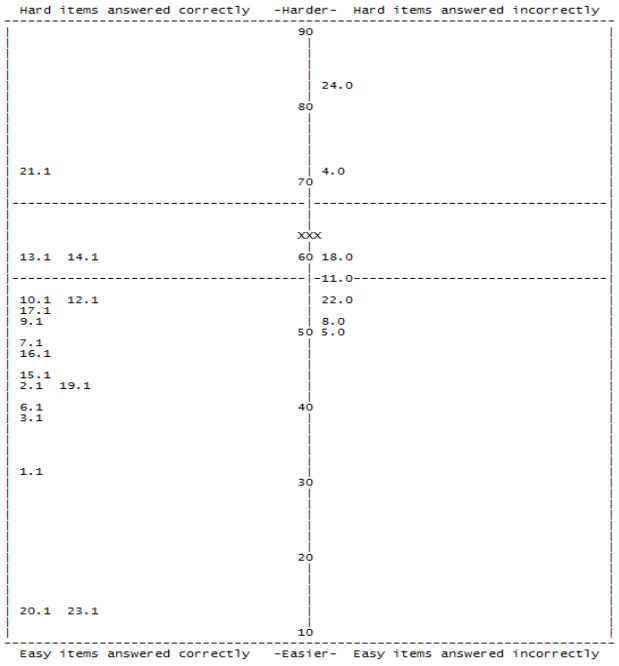

Figure 1. A simple reporting model of student ability against items and class average

For example, you could adopt a model (Figure 1) that reports, *'this is where the child is'*, subject to the margin for measurement error. Here are the items correct; anything correct above the ability line can be regarded as a strength and unexpected errors below the ability line can be considered a weakness. The class average can be similarly located on the same assessment graph. The research of Prof Yan (2022) reports on developing a vertical scale for mathematics, (which is what is needed to implement this across a primary school system), and then have applications in classroom testing.

In all, that's a very modest proposal about what can be done. It would have been a less modest proposal if the progress already made in Hanoi with adaptive testing was obvious (Nguyen et al., 2022). But, there's a problem that remains for even though Hanoi has a quality academic research group who can develop a testing system at a university, actually being able to get permission and collaboration with children and teachers in the classroom to use the system, is not a straightforward matter.

6. Summary

Formative assessment, not summative assessment; assessment for learning, instead of assessment of learning, find where the pupil is, see the gaps, make them grow. Track their growth. What's the most important thing? Get teachers, parents and students on side. Rasch measurement is not the answer. Assessment for learning is the answer. And Rasch measurement is the technique by which we can implement that. You need to copy the practices of others, but, don't copy NAPLAN. Copy e-asTTle. Copy what's being done in NWEA. You won't get there straight away. It costs a lot of money, but start now by item banking maths and science, and letting teachers use it.

Notes

Details of the testing systems can be accessed as follows:

The National Assessment Program - Literacy and Numeracy (NAPLAN) https://www.nap.edu.au/home

e-asTTle features https://e-asttle.tki.org.nz/About-e-asTTle/Features

MAP Growth https://www.nwea.org/map-growth/

References

Brown, G. T. L. (2013). asTTle-A national testing system for formative assessment: How the national testing policy ended up helping schools and teachers. In M. K. Lai & S. Kushner (Eds.), *A National Developmental and Negotiated Approach to School and Curriculum Evaluation* (pp. 39-56). Emerald Group Publishing. https://doi.org/10.1108/S1474-7863(2013)0000014003

Brown, G. T. L., O'Leary, T. M., & Hattie, J. A. C. (2018). Effective Reporting for Formative Assessment: the asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 107-125). Routledge. https://doi.org/10.4324/9781351136501-11

Dang, X. C. (2022). *The presentation of assessment in the new general education curriculum in Vietnam: A proposal for developing instruments using Rasch measurement.* In Pacific Rim Objective Measurement Symposium, Hanoi, Vietnam.

Ingebo, G. S. (1997). *Probability in the measure of achievement*. Chicago, IL: MESA Press.

Nguyen, T. H., Le, T. H., & Tran, X. C., (2022). *Assessment of mathematical modeling competency grade 12 students using computerized adaptive test*. Paper presented at the Pacific Rim Objective Measurement Symposium, Hanoi, December.

Wilson, K.M. (1988). A Study of the Long-Term Stability of GRE General Test Scores. *Research in Higher Education,* 29(1), 3-40.

Yan, Z. (2022). *The presentation of applications of Rasch measurement in research and practice.* In Pacific Rim Objective Measurement Symposium, Hanoi, Vietnam.