BEST CASE SCENARIO ASSESSMENT

INTRO

I must admit I am quite impressed that the HTAV was willing to have not just one article, but an entire episode of this professional publication on the topic of assessment. Considering how poorly assessment has been used in schools for decades, with confused purposes that contradict each other, and outlandish increases in teacher workload which often amount to very little improvement in student outcomes, it is no surprise that the topic is one that few teachers enjoy. Indeed, any time I speak on the topic and find that the room is not empty, I'm pleasantly surprised.

A few comments about terminology are warranted first up, I believe.

JARGON

You will often hear in education of the difference between formative assessment and summative assessment. The often-quoted analogy runs like this: formative assessment is when the chef tastes the soup, summative assessment is when the customer tastes the soup. The point being, formative assessment can change the soup (read: student outcome), whereas summative assessment generally doesn't. Firstly, the analogy is bogus: if all your customers complained about how gross the soup was, might it not be changed? Secondly, assessments themselves are not formative or summative, it is the way that we use the information that can be one or the other. Imagine a medieval history topic test. If I set this in Week 1 of Term 1 and use that information to leave out bits of the curriculum most students already know about, and also to create ability or mixed-ability groups for learning activities, I am using the information in a formative way – it forms (or influences) my future teaching. If, however, has is far more often the case, I set a topic test at the very end of a unit, feed the resulting number into my school's database and move on to another topic, I am using the information summatively – it sums up how much a student knows about a topic. Let's put aside the not insignificant issue of measurement error for a moment, and assume a test can give me a summation of a student's learning up until that point. The reason for this formative/summative dichotomy I believe is to point out that, in my opinion, one is useful and the other is not. Formative use of assessment data can be used to improve your teaching. Don't teach students what they already know. Don't teach them things that are not developmentally appropriate for them. Formative use of assessment information is the first step to differentiating teaching based on ability. Summative use of assessment information is for things like school reports and ATARs. Do school reports and ATARs improve learning? In one negative sense they might. Beside they are so high stakes, students might put in more effort to get a good report or ATAR. However, there are a multitude of local examples of the ATAR system, for example, causing a lot of student anxiety and stress, at a very delicate time in their lives (see, for example, Bowden, 2017; Corrigan, 2020; Fischetti et al., 2020).

The purpose of assessment in education has been murky for a while now. Perhaps in some bygone era assessment was simply used to weed out the less academic types who would go on to a lifetime of fulfilling skilled or semi-skilled labour, leaving limited places in senior secondary for those destined for the professions.

In other industries, assessment is used as information. How many cars is my factory producing? X per day. How can I increase that number? Assess the variables and try and improve in one of them. Assessment in these situations is used as information with which to make decisions to improve a process. Unfortunately, for too long assessment has been used as judgement: judgement of students, teachers and schools (and most of it negative). Assessment has a justifiably bad reputation in education because its use is seen as disparaging, and it is rarely used to improve processes (except for lip service).

BIG DATA VS SMALL DATA

There is a lot of buzzing about 'big data' in the corporate world, and as managerialism (the idea that professional managers should run things, not specific experts) has begun to take over the education system, the spectre of big data has fallen onto schools. But big data is almost always irrelevant to the classroom teacher. Don't worry about it. Even NAPLAN results cannot be used to infer student growth, as the measurement error is larger than the expected growth (Wu & Hornsby, 2014), shocking yet little known fact. What is important is small data, data that you generate yourself. Assessment data should, in its best sense, only be generated when you have already thought of something you want to do with it. Unfortunately, the government requires schools to rate students on a minimum five-point scale at least twice per year ("Australian Education Act," 2013). This is the ultimate kind of summative assessment data – it isn't really used to improve student outcomes or teaching. How could it? If a student gets a "A" or a "80%" on an assessment, what information does it give teachers or learners? Get an A+ next time? Try and get the remaining 20%?

We have two kinds of assessment data – qualitative and quantitative. Put simply, quantitative might be "80%". Qualitative might be "is able to multiply numbers up to 100". For similar reasons just discussed, qualitative assessment information is a lot more useful than quantitative. Why is there still so much quantitative data around, you might ask? I think it has to do with the fact that it is easier to gather and is what has been done in the past. Neither of which are good reasons to continue doing it into the future.

NORM-REFERENCING VS CRITERION-REFERENCING

A final note about vocabulary. There are a couple of common ways to assess learners – against each other (norm-referencing) or against set standards (standards referencing or criterion referencing). I'm hoping by now you can figure out which one I prefer. Are we going to improve the learning of students by pitting them against each other, or by showing them what they are capable of doing? Most curricula, and the outdated ATAR system use norm-referencing. Those bland capabilities you read about in the Victorian Curriculum are norm-referenced, stating basically that if the learner was manufactured (read: born), in say, 2006, they are now 15 years of age and therefore 'should' be at the standard listed in whatever it says for the Year 10 curriculum. Of course, as anyone who has taught for five seconds knows, these 'norms' are only true of about the middle 30% of any class. In a typical Year 10 class the expected range of ability can spread from Grade 6 all the way up to second year university (Smith, 2015). What exactly is the point of having the same expectations of all these various students?

WHATS THE PURPOSE?

So, what is the point of assessing students? An often-heard trope again assessment is that it doesn't matter how many times you weigh a pig, it doesn't get any fatter. True that. One would hopefully assume that there is some fattening up of the pig going on in between trips to the scales. There's a couple of questions we need to ask ourselves first. Are we going to be able to meaningfully interpret the results of our assessment? If a student gets a "B" – what does that *mean?* How can I decode the information contained in the "B"? Secondly, is the assessment going to improve student learning, or, to put it another (but hopefully similar) way is, is the assessment going to improve how I teach that student? If the answer to both of those questions is no – No, I won't know how to meaningfully interpret it or no, the information won't be used to improve teaching and learning – what is the point? Reporting on current achievement levels to parents and the government is obviously one common answer. However, as suggested, this isn't a great use of student and teacher time. A related question is: do you already know the answer you're seeking? In many instances, students and teachers know in advance what the results of an assessment are going to be. Therefore again, what is the point? One of the highest effect sizes Hattie (2009) has uncovered is student's saying what grade they are going to get on an assessment before they even do it. In such instances, save time and just get the student (or teacher) to judge in advance where they are at. A classic case in point: as my old boss Patrick Griffin used to say – NAPLAN

results have largely flatlined for 10 years so what is the point of spending hundreds of millions of dollars and losing a week of teaching to find that out?

DIFFERENTIATION

Which brings me to what I would argue is the best use of assessment information: to diagnose a student's 'point of readiness' or Zone of Proximal Development (Vygotsky, 1962) or even catchier, the student's 'goldilocks zone' (not too hard, not too easy). Once you know that, teach them from that point. Some research suggests up to 80% of what comes across the average secondary school student's desk is either too easy or too hard. 80%!

There are some obvious criticisms of targeting instruction at point of readiness, aka 'differentiation by ability'. I think there is one major one that deserves our attention – it is just too hard and/or time-consuming (Ashman, 2020). I completely agree, assuming that teachers work by themselves. But as many of us know, 'Collective teacher efficacy' is literally the factor with the highest effect size of them all (Hattie, 2009). Here is a suggestion for how a teaching team of five could plan a differentiated curriculum (for both content and skills) for a 10 week term of history, assuming 3 lessons per week:

Assumptions:

- You're already using a textbook (which qualifies as the mid-level literacy document)
- You're teaching five historical skills
- You've mapped out five levels of sophistication for each historical skill
- Which means 5x5=25 quality criteria
- It takes about 30 minutes to write a targeted activity, of the kind I've written (which you can have for free on my website) (Lawless, 2020)
- The content knowledge written at a lower literacy level you could either rewrite the textbook chapter (time-consuming) or find something online (quick and easy)
  - 5 minutes per piece of knowledge
- The content knowledge written at a higher literacy level, ideally you would just find a more complex article about the topic from a free online encyclopaedia (even Wikipedia provides great in-depth and expert-vetted materials) or another online authoritative source
  - 5 minutes per piece of knowledge

Therefore you need:

- 20 targeted activities (not 25 because, according to the rubric writing guidelines I adhere to and you should too, the lowest level skill should be something everyone in your class can already do)
- 30 pieces of content knowledge written (or more likely, sourced) at a lower literacy level
- 30 pieces of content knowledge written (or more likely, sourced) at a higher literacy level

Task 1	Task 1 time cost	Task 2	Task 2 time cost	Task 3	Task 3 time cost	TOTAL
Write 2 targeted activities	1 hour	Source 3 pieces of content knowledge at lower literacy level	10 minutes	Source 3 pieces of content knowledge at a higher literacy level	10 minutes	1 hour 20 minutes

Thus, for just 1 hour 20 minutes per teacher, in two years you would have a fully differentiated curriculum, with 3 levels of literacy and 5 levels of skills covered. If you increase the time teachers spend on shared planning for differentiation by, I don't know, say using one of the many 'assessment' or 'PD' days the union keeps winning, you could do an entire year's worth of differentiated planning in a single day. I am flabbergasted that this is not a common practice. My attempt to make a bank of targeted differentiated activities can be found here: https://lawlesslearning.com/free/compositerubrics/

PRE- AND POST-TESTING

Another assessment practice that is lacking or non-existent in the history space is pre- and post-testing. Content pre and post-tests can be useful to help teachers find out what students didn't learn, and therefore allow them to do some re-teaching if needed. Perhaps even more useful for improving teaching effectiveness is content-agnostic skills pre- and post-testing. That is to say, you create a test that assesses student history skills (e.g., the big five from the Victorian Curriculum) which doesn't require any content knowledge, or at least, the question itself has all the knowledge needed to answer it. You pre-test at the start of the history teaching part of the year and the post-test with the same test at the end of the history teaching part of the year. You can generate a number, 'effect size', which tells you how much impact instruction has had. One of Hattie's (2009) mantras is, "know thy impact". With this system, you can find out which teachers are the most effective and see whether anything they're doing is transferrable to other teachers. A really important message, and something I think that holds our profession back somewhat, is that teaching is a skill, and some people are better at it than others. We can't hide from this fact. Just as we put our learners through a cycle of performance, judgement, feedback and improvement, we should be courageous and confident enough to expect it for ourselves. Effect sizes are perhaps most powerful in a school context at comparing year level courses. If your Year 7 results show that students developed a lot in their source analysis, but Year 8 students did not, that could be a good opportunity to ask questions about the two courses. What is it in the Year 7 course that helps student develop that skill?

How effect sizes are calculated is explained here: https://www.youtube.com/watch?v=iMGLzASbjYI A spreadsheet that will do the calculations can be found here: https://evidenceforlearning.org.au/assets/Evidence-Exchange/Effect-Size-Calculator-Excel.xlsx

FINAL WORD

There is a lot to say about assessment, more than can be discussed here. Let me just leave you with these thoughts, albeit my own opinions:

1. Formative *use* of assessment has a much bigger impact on learning than a summative use of assessment information
2. Assessment should be used as information to improve teaching or make other education-relevant decisions, not to needlessly judge students
3. Teachers should be detectives and use 'small data' assessment to find things out that they need to know in order to improve a student's learning
4. Assessment should be about what students can do, not how they compare to others. Perhaps there are instances in society when comparison is relevant, but in most schooling situations, these comparisons are just demotivating and stressful
5. Use assessment to target instruction
6. Differentiation is not possible without shared planning
7. Pre- and post-testing of skills can be used to determine teacher and, more usefully, course effectiveness

REFERENCES

Ashman, G. (2020). *The power of explicit teaching and direct instruction*. Great Britain: Corwin.

Australian Education Act, (2013).

Bowden, T. (2017, October 11). Year 12 exams — are they worth the stress? ABC News. https://www.abc.net.au/news/2017-10-11/year-12-exams-are-they-worth-the-stress/9029260

Corrigan, De. (2020, April 2). To scrap or not to scrap: The bigger question beyond the ATAR. Monash Lens. https://lens.monash.edu/@education/2020/04/02/1379964/covid-19-and-year-12-anxiety-should-the-atar-be-scrapped

Fischetti, J., Smith, M., & Varanasi, R. (2020, July 27). We know by Year 11 what mark students will get in Year 12. Do we still need a stressful exam? The Conversation. https://theconversation.com/we-know-by-year-11-what-mark-students-will-get-in-year-12-do-we-still-need-a-stressful-exam-140746

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Lawless, B. (2020). Composite Rubric – History. https://lawlesslearningland.files.wordpress.com/2020/03/composite-rubric-history.pdf Accessed on 29 March 2021

Smith, A. (2015, July 26). Grattan Institute numbers show teachers face wide range in ability within class. The Sydney Morning Herald. https://www.smh.com.au/education/grattan-institute-numbers-show-teachers-face-wide-range-in-ability-within-class-20150725-gikakj.html

Vygotsky, L. (1962). Thought and language (E. Hanf-mann & G. Vakar, Trans.). Cambridge, MA: MIT Press.

Wu, Margaret; Hornsby, David (August 2014). "Inappropriate uses of NAPLAN results". Practically Primary. 19 (2): 16.