Interpreting learning progress using assessment scores: What is there to gain?

Dr Nathan Zoanetti
Australian Council for Educational Research

https://doi.org/10.37517/978-1-74286-638-3_17

Dr Nathan Zoanetti joined ACER in 2018 as Research Director of the Psychometrics and Methodology Program and before that he was Principal Psychometrician at the Victorian Curriculum and Assessment Authority. He has previously held research positions at the Assessment Research Centre, University of Melbourne, including as Executive Officer of the NAP Literacy and Numeracy (NAPLAN) Measurement Advisory Group, and more recently as an Honorary Senior Fellow. He is currently a member of the OECD PISA 2021 Creative Thinking Expert Group and a member of England's Standards and Testing Agency Technical Advisory Group.

Abstract

Using assessment scores to quantify gains and growth trajectories for individuals and groups can provide a valuable lens on learning progress for all students. This paper summarises some commonly observed patterns of progress and illustrates these using data from ACER's Progressive Achievement Test (PAT) assessments. While growth trajectory measurement requires scores for the same individuals over at least three but preferably more occasions, scores from only two occasions are naturally more readily available. The difference between two successive scores is usually referred to as gain. Some common approaches and pitfalls when interpreting individual student gain data are illustrated. It is concluded that pairs of consecutive scores are best considered as part of a longer-term trajectory of learning progress, and that caveated gain information might at best play a peripheral role until additional scores are available for individuals. This review is part of a larger program of research to inform future reporting developments at ACER.

Introduction

Progress can be quantified using assessment scores as soon as two score points are available for the same individual. However, there are well-known technical shortcomings associated with quantifying progress based on only two scores (Willett, 1994; McCaffrey et al., 2015). These limitations stem from unavoidable causes including natural variation between students' rates of learning progress, and margins of error associated with the assessment scores themselves (Singer & Willett, 2003). Failure to account for these factors can result in spurious classifications and comparisons of progress for a non-trivial proportion of students. This paper argues that placing too much emphasis on individual progress metrics that are based on only two scores is likely to be counterproductive in practice. Instead, it is concluded that pairs of consecutive scores are best considered as part of a longer-term trajectory of scores along a clear progression of learning.

Defining gain and growth

Recommendation 4 of *The report of the review to achieve educational excellence in Australian schools* (Department of Education and Training, 2018) draws attention to the importance of gain and growth and invites clarification of the definitions of these terms:

> Introduce new reporting arrangements with a focus on both learning attainment and learning gain, to provide meaningful information to students and their parents and carers about individual achievement and learning growth (p. xiii).

Noting that terminology about learning progress can be varied (Hollingsworth et al., 2019), in this section we refer primarily to references that are concerned with quantification of progress using assessment scores. Assessment scores in isolation are sometimes called status measures (Castellano & Ho, 2013a). Terms like achievement and attainment are also used, as seen in the above recommendation. Moving beyond status to consider progress, it is generally accepted that progress measures require scores from the same student or students on multiple occasions. These serial data are referred to as longitudinal.

Contemporary research and practice on reporting progress using assessment scores reveals that many implementations are limited to quantifying progress using scores from only two successive occasions (O'Malley et al., 2011). Nese et al. (2013) and Ployhart and MacKenzie (2015) point out that this 'change score' between two occasions does not properly characterise growth, but instead would be more accurately characterised as gain. This seems like a useful distinction given the increased complexity of the statistical models that accommodate scores from more than two occasions and the more robust inferences about progress they can support (Curran et al., 2010).

The technical superiority of growth measures has at least two contributing factors. First, with the additional data points it is possible to average out or statistically account for measurement error and other statistical artefacts that plague simpler gain measures. Second, there is the capacity to construct and compare trajectories that contain nuanced information about growth by modelling change over time as a continuous process (Willett, 1994). Nonetheless, the naturally greater availability of gain information relative to growth trajectory information provides strong motivation to make use of the former whilst accommodating its limitations.

Preconditions for meaningful progress measurement

For gain or growth modelling that can meaningfully be related to learning in a given domain, the assessment should ideally have the following characteristics:

- all scale scores within a domain within the same assessment program should be on a common 'vertical scale' with interval properties
- each assessment already has, as part of its reporting framework, described proficiency levels that provide a criterion-referenced interpretation of progress.

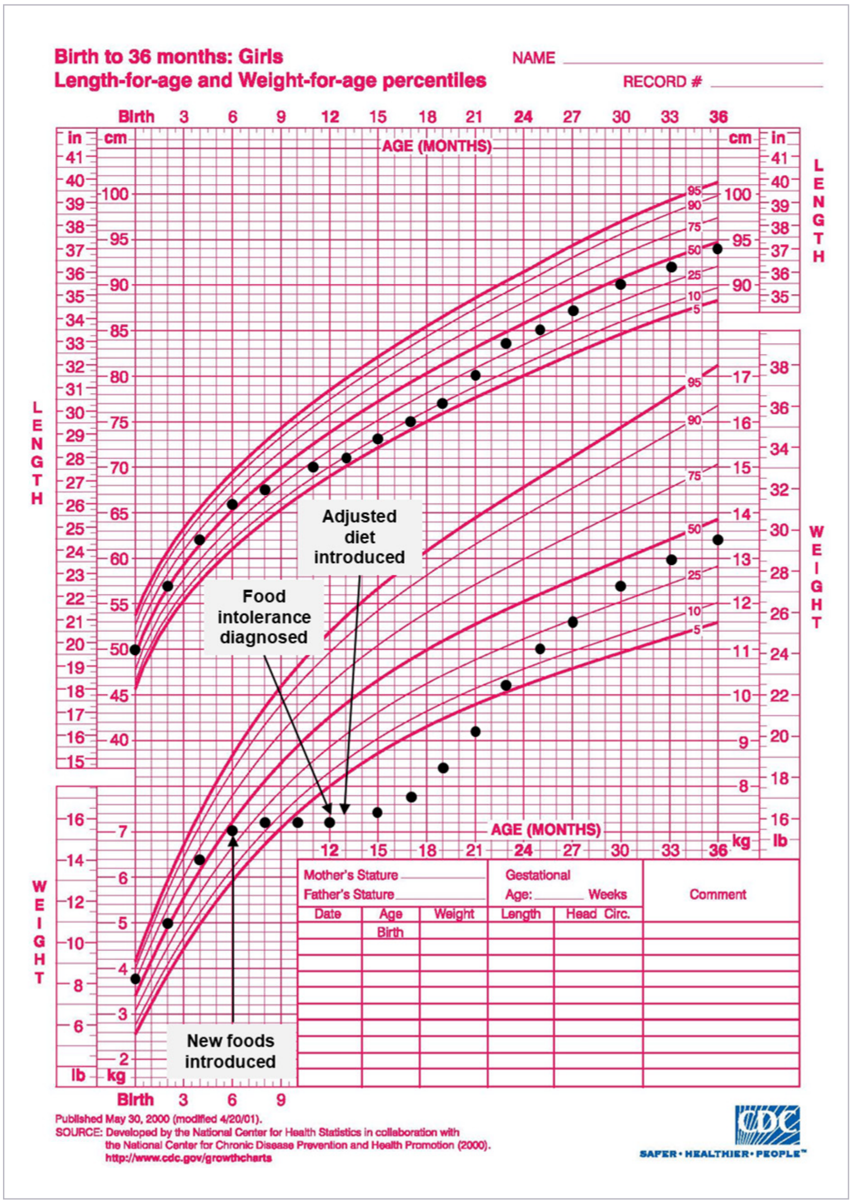
If these conditions cannot be met in practice, then there will be limitations on the range of methodological options for modelling and interpreting progress in a valid and meaningful way (Patz, 2007; Protopapas et al., 2016; Sireci et al., 2016).

What does learning growth typically look like on a scale?

This section summarises what learning growth often looks like when evidenced using assessment scale scores. This provides an important basis for contextualising changes in scale scores from one occasion to the next, particularly as they relate to making quantitative comparisons between the progress of individuals and groups. First though, it is instructive to consider growth against well-defined scales that measure attributes other than learning.

A well-known example comes from paediatric contexts, where measures such as the height, weight and head circumference of infants over time provide key developmental indices (see Figure 1). Much as in learning contexts, substantial deviations from typical trajectories can indicate that additional or different interventions are required. In these cases, the trajectory of growth following the intervention becomes of central interest. Parallels can be drawn with education, though successive measures from educational assessments are typically much more variable.

Figure 1 Example of a child's weight trajectory recorded on a paediatric growth chart



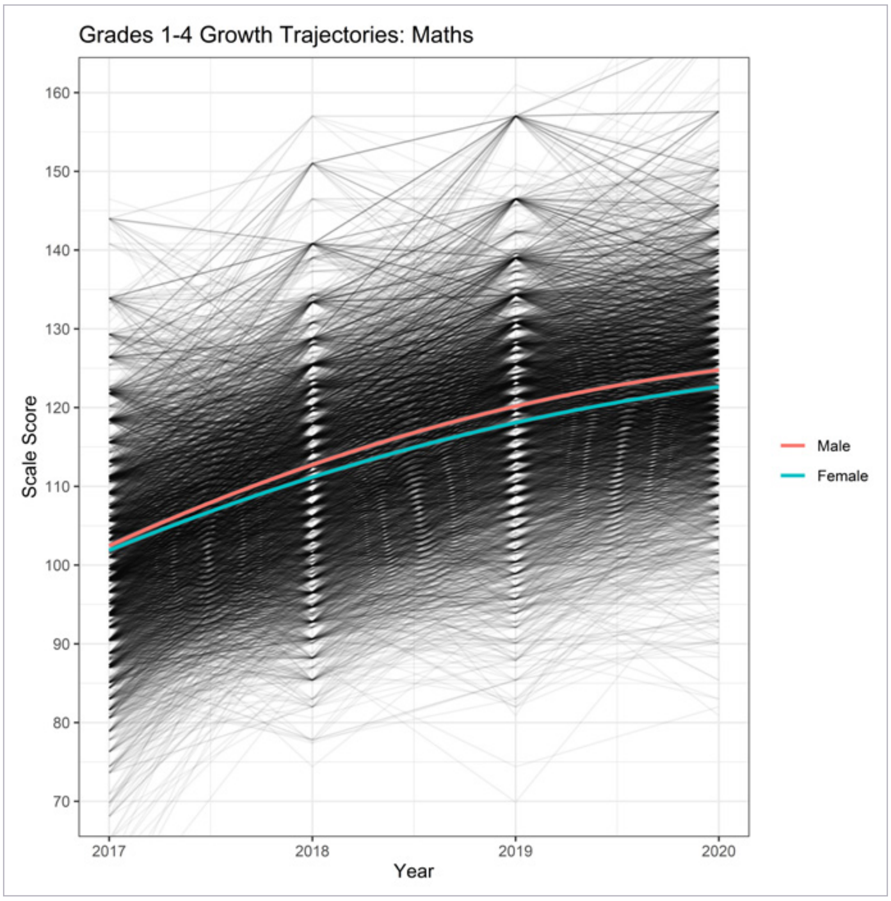
Growth chart template sourced from Kuczmarski et al. (2002, p. 36)

For many human attributes, evidence suggests that it is common for more substantial gains to be made initially. Growth rates often decelerate or stabilise with increasing amounts of the attribute. A brief scan of standardised assessment results and research literature from developmental psychology and school education contexts suggests that similar cohort-level patterns are commonplace (Australian Curriculum, Assessment and Reporting Authority, 2019; Li-Grining et al., 2010; Morgan et al., 2009; Williamson, 2018). However, this is not necessarily the case for all domains and age groups (Castellano & Ho, 2013a), and it is seldom true of every individual's growth trajectory.

Also of interest in educational research and evaluation is whether the growth trajectories of different groups of students differ. These groups may be categorised by contextual variables (e.g. school type), student characteristics (e.g. gender) or initial achievement levels (Singer and Willett, 2003). Whether the growth trajectories of different groups converge or diverge is also of key interest for detecting the so-called Mathew effect (Merton, 1968). In education this effect manifests as achievement gaps between groups that increase over time (Pfost et al., 2014).

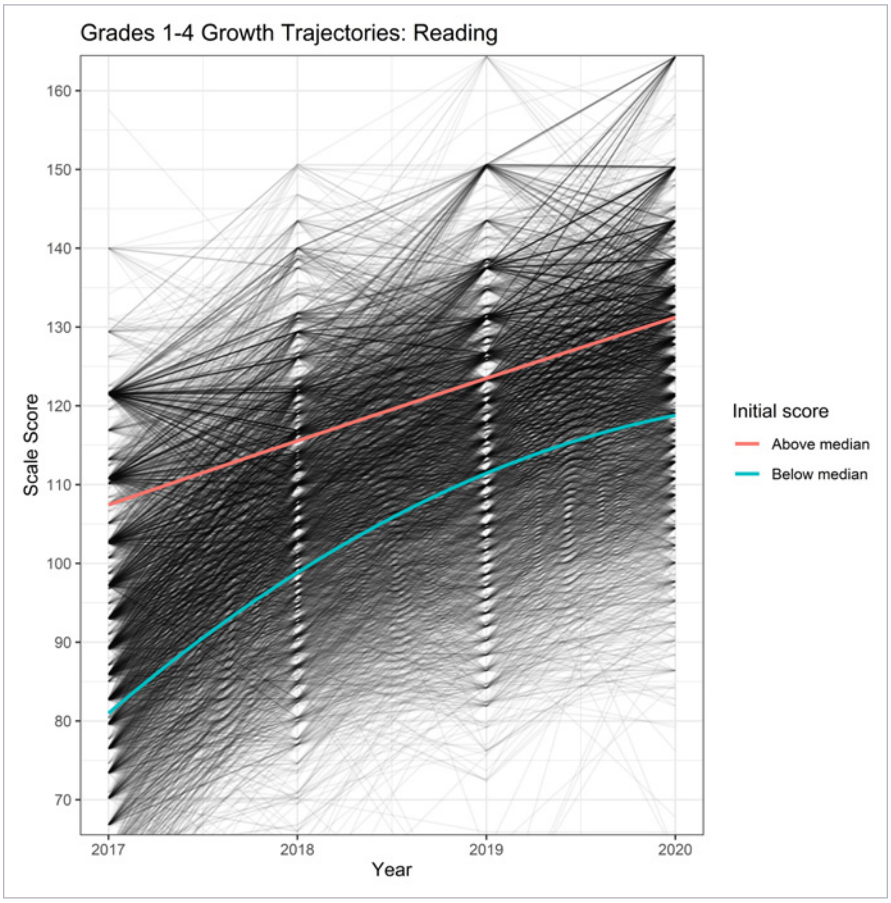
Figure 2 shows the average growth trajectories of female students and male students in a large sample of longitudinal Grades 1–4 PAT Mathematics data from Term 4 sessions. The grey lines in Figure 2 (sometimes referred to as a spaghetti plot) show how varied and volatile the observed initial scores and score gains can be for individual students. While it is difficult to visually discern, the volatility is greater among students with relatively extreme scores. After applying a three-level random intercept mixed-effects regression model using the lme4 (Bates et al., 2015) package in R (R Core Team, 2020), the following model parameterisation was well-supported: a quadratic (i.e. curvilinear) growth model fitted the data better than a linear growth model ($\chi2(1, N = 20776) = 753.1$, $p = .00$); and, consistent with Figure 2, allowing the slope but not the intercept to differ across female and male students yielded the best model among several that were compared.

Figure 2 Individual and average growth trajectories for Years 1–4 PAT Maths for female and male students



The example in Figure 3 from PAT Reading shows initial convergence in the trajectories of students grouped by starting score (above or below the median) followed by more consistent growth rates, albeit at different levels. Looking at only the first year of progress, it is natural to conclude that the lower achieving group is making rapid progress in their learning and is on track to bridge the achievement gap, but there is a catch. The initially pronounced convergence observed here is in part a statistical artefact of having selected these groups on the basis of their initial scores. Grouping students in this way introduces an upward bias in the low scoring group, and vice versa, by inducing what is referred to as regression to the mean (Barnett et al., 2005). This phenomenon can have profound implications for interpreting scale score gains and is outlined in more detail in the following section.

Figure 3 Reading growth trajectories for students grouped by initial score above and below the sample median



Interpreting individual student gains

The discussion and patterns reviewed so far suggest the following tendencies:

- there is considerable variation in initial scores
- score gains can be volatile, particularly for students with extreme scores
- score gains sometimes taper off as students progress further up the scale
- score gains may differ between students grouped by certain characteristics.

The interpretation challenge is to take these observations into account when appraising an individual student's gain from one assessment to the next, and when making comparisons between the gains made by different students. The following two factors contribute to the observations just listed and have direct implications for interpreting gains:

- rates of learning (actual progress) vary across individuals and groups
- scores from all assessments contain measurement error.

These two factors introduce natural variation in the scale scores attained by students over consecutive occasions. This results in an imperfect level of correlation between initial scores and final scores, and imperfect correlations will always be accompanied by regression to the mean (Kahneman, 2011).

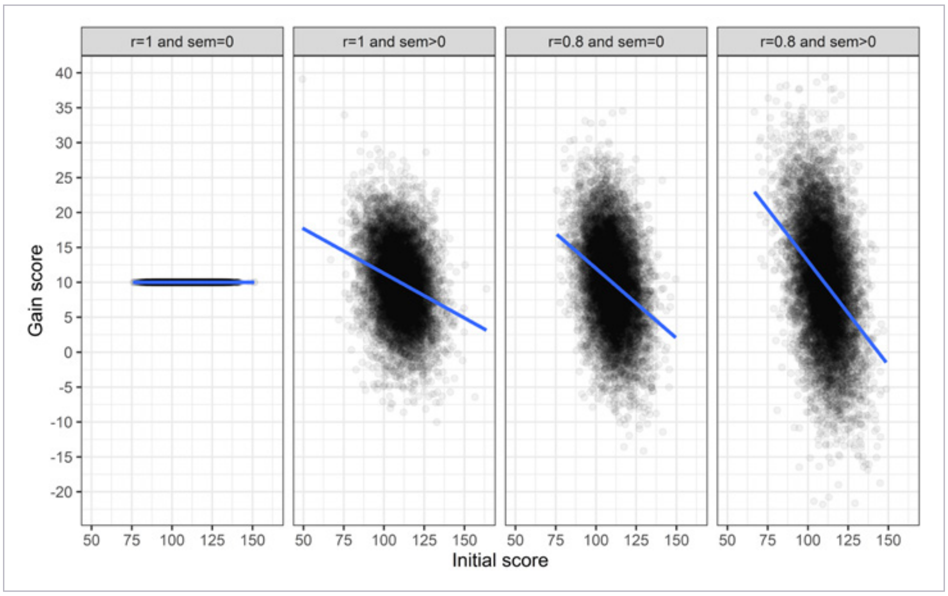
If on one measurement occasion random or idiosyncratic variation has a relatively large impact on a student's scale score, it becomes likely that on the second occasion it will contribute less. These statistical artefacts, particularly in a learning context characterised by genuine decelerating growth, produce the '…well-known negative correlation between prior score and gain …' (Betebenner & Linn, 2009, p. 6). In particular, students with prior scores higher than the population mean will systematically tend to show lower gains, and vice versa. Regression to the mean can make natural variation in repeated data look like real change.

Figure 4 illustrates this phenomenon by comparing the relationship between gain scores (i.e. final score minus initial score) and initial scores under different simulated conditions with the following parameters:

- population size of 10 000 students
- initial scale scores with a mean of 110 and a standard deviation of 10
- final scale scores with a mean of 120 and a standard deviation of 10
- latent or true correlation ('r') between initial and final scores of either 1 (i.e. all students gain exactly 10 scale scores) or 0.8 (close to that for PAT assessments taken one year apart after disattenuating for measurement error)
- measurement error ('sem': standard error of measurement) set at either zero (perfectly precise measurement) or between 3.5 and 6.5 following a quadratic error function giving extreme scores larger errors (errors are assumed to be uncorrelated).

The blue Loess fit lines in Figure 4 provide a moving average of the gain scores across the initial score range. These show that regression to the mean occurs as soon as there is measurement error in the assessment or as soon as there is an imperfect level of correlation between initial and final status. It is also clear that these two factors have a cumulative impact. Comparisons of score gains with the average would be biased between 0–3 scale score points across the middle 95 per cent of initial scores in the most realistic scenario (right-hand panel). The direction (positive or negative) is determined by whether the initial score was above or below the population mean of 110. Larger systematic bias is present as expected for students with more extreme initial scores. Comparing gains between students with initial scores either side of this range will be subject to biases exceeding half of the average population gain made in one year.

Figure 4 Four simulated gain scenarios illustrating regression-to-the-mean in absolute gain measures



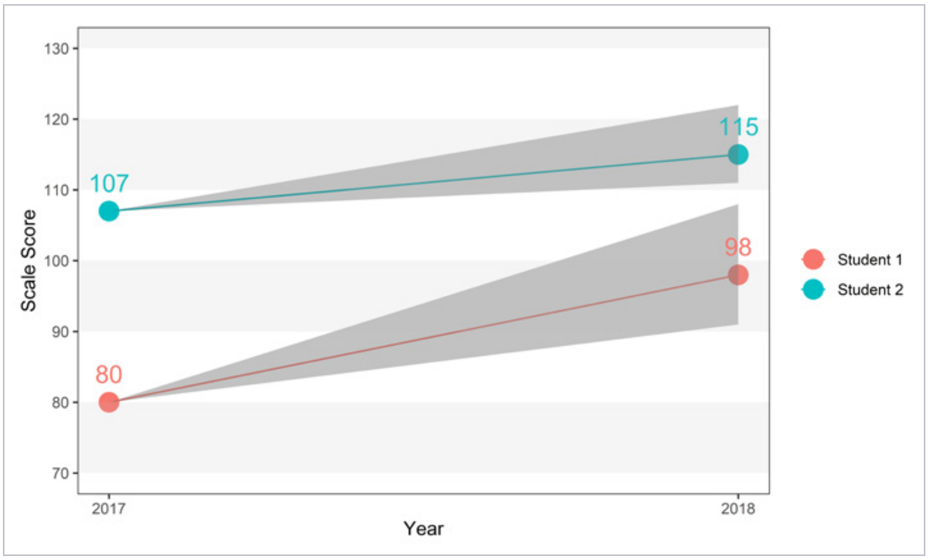
The simulation here has focused on revealing one unavoidable source of bias that can impact gain comparisons between students with different initial scores. Also worth noting is the unavoidable variation in gain scores resulting from the presence of realistic levels of measurement error. This can be seen clearly by examining the differences between the gain values in the first and second panels in Figure 4. The vertical spread of gain scores observed in the second panel but absent in the first panel is entirely attributable to measurement error. Depending on factors like the time between assessments, and the targeting of the assessments, this variation may be substantial enough to mask true gains or to mask biases due to regression to the mean. This is relevant for a non-trivial proportion of students, some of whom would attain a negative gain value due to chance alone. This matters when the focus is on quantifying, appraising and communicating individual student gains. Some researchers have argued that gain measures can be reliable when score distributions have certain characteristics (e.g. Rogosa et al., 1982; Williams and Zimmerman, 1996). However, we have observed that these characteristics are unlikely to apply to scale scores from high-quality, well-targeted assessments taken one year apart. The latter tend to more closely approximate distributional and correlational conditions known to be associated with low gain score reliability (e.g. Cronbach and Furby, 1970).

Statistical corrections can be made to account for regression to the mean in some situations (Rogosa et al., 1982), but this is not always practical or technically feasible. Therefore, in addition to expecting some volatility in absolute gain measures, anticipating asymmetries in gain scores across the initial score range is critical for ensuring that changes in scores are responded to proportionately. This in turn ensures that learners and educators are supported to direct their efforts in a targeted way. Being able to avoid incorrect conclusions about student progress, such as that a school appears to be doing a better job improving the learning of its lower achieving students than its higher achieving students according to gain scores alone, is one example of why these statistical considerations matter in practice.

A consequence of these biases is that absolute gains are often perceived as unfair for comparing the progress of individuals and groups who differ substantially in their prior achievement. To help contextualise whether an observed absolute gain is 'typical' or otherwise in the presence of these biases, it can be helpful to draw upon normative information. Several norm-referenced interpretations are possible, starting with simple comparisons to available cross-sectional scale score norms (like in Figure 1) and progressing to conditional metrics that take into account prior achievement and possibly contextual variables.

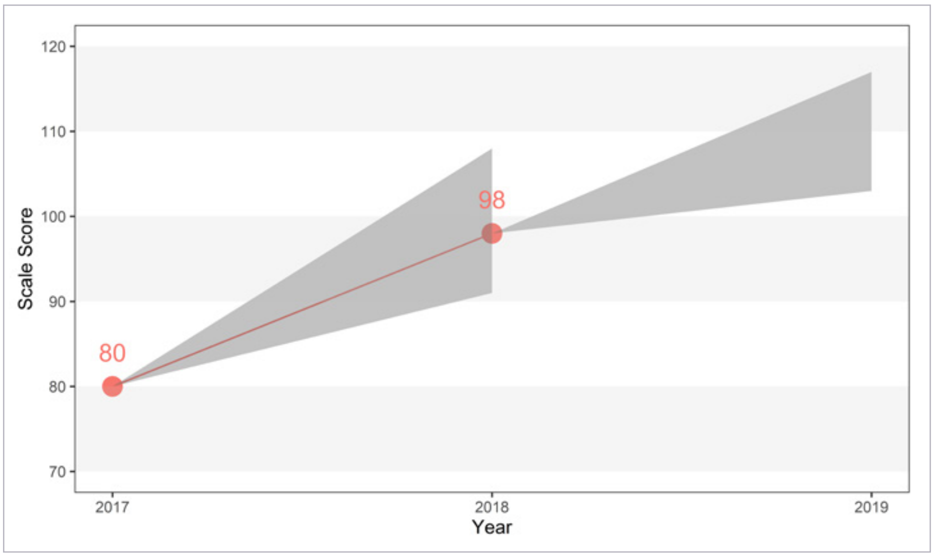
The use of models that compare progress between students with similar prior scores has emerged as a popular way to make these biases less visible through the construction of ostensibly fairer comparison groups for each individual student. A simplified version of this approach based on calculating absolute gain percentiles for students grouped by similar prior scores is shown in Figure 5. The middle 50% of these relative gain percentiles is shaded dark grey. It is worth noting that a variety of alternative calculation methods exist, including relative gain or conditional status measures (e.g. Castellano & Ho, 2013a, 2013b) and Student Growth Percentiles (SGPs) (Betebenner, 2011). Here we will use the term relative gain percentile since we adopt a simplified percentile-based calculation rather than a conditional regression-based calculation. It can be seen that the different levels of absolute gain for Student 1 (18 scale scores) and Student 2 (8 scale scores) both result in relative gain percentiles that are close to the middle of their respective relative gain distributions. The grouping by prior scores has ameliorated some of the biases that undermine comparisons between students who start at markedly different locations on the scale.

Figure 5 Comparison of gains for students with markedly different initial scores



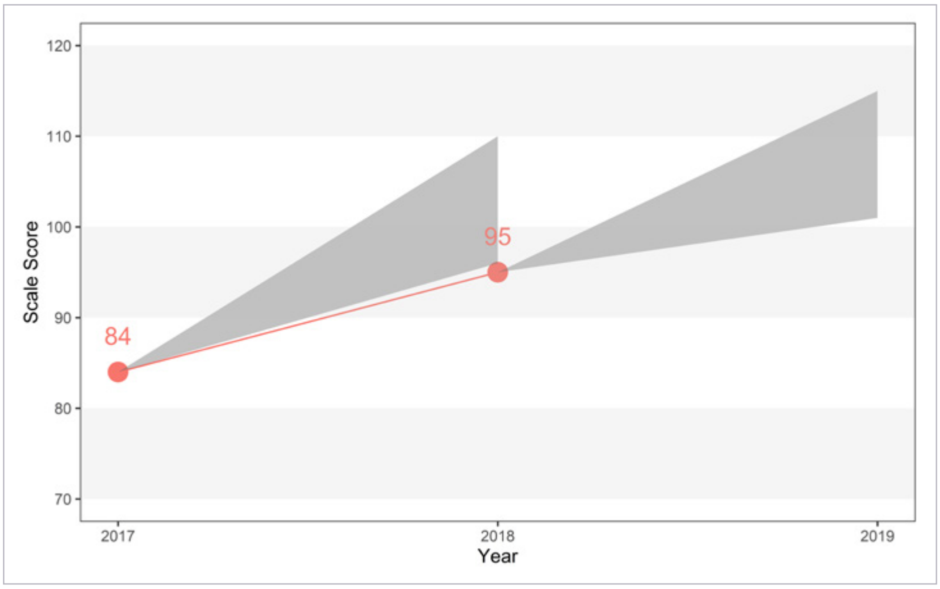
Using recent historical data from the same assessment, it is also possible to show projections of typical gain ranges that take prior score into account. Projections like these can be found in some reporting systems (Betebenner, 2011). The projection in Figure 6 shows the range of scale scores obtained historically by the middle 50 per cent of students who had also started with a scale score close to 98 one year prior. This kind of depiction may be useful for stimulating discussion and setting expectations about future learning goals and progress.

Figure 6 Relative gain percentile distribution for recent scores and as a projection



Unfortunately, the metrics that these relative or conditional models produce, if based on only one prior score, are volatile (McCaffrey et al., 2015; Sireci et al., 2016). For Student 1 and Student 2 in the earlier example, whose relative gains placed them close to the median, they could with non-trivial probability be classified as being in the lower or upper relative gain quartile after allowing for a realistic perturbation of scale scores by approximately one standard error of measurement (usually 3 to 4 scale score units). This is illustrated in Figure 7.

Figure 7 Illustration of classification volatility of relative gain percentiles



This simplified example is emblematic of the non-trivial levels of misclassification that can arise when using relative gain or conditional status metrics for individual students. From a measurement standpoint, there is little impact on the substantive interpretation of knowledge and skill for scale scores that have been perturbed within the bounds of measurement error. Correspondingly, the achievement bands would be stable or at most would change by one band near the level boundaries. In contrast, it is not unusual for standard errors associated with relative gain or conditional status percentiles to be as large as 15 percentile points (Sireci et al., 2016). In this situation, an estimated one-in-ten students with an observed relative gain percentile of 50 could be operating in the upper or lower conditional gain quartiles. This is consistent with modelling by Betebenner et al. (2016) who showed that approximately one-in-six students with an observed conditional percentile of 50 might in reality be below the 35th percentile progress benchmark used in that context. It follows that caution is required when interpreting individual student gain metrics like these and when using them to label the gains of individual students as typical or otherwise.

While conditional or relative gain approaches largely overcome comparability biases due to regression-to-the-mean and tapering growth trajectories, their apparent accentuation of measurement error is an unfortunate shortcoming. This brings into question the reliability of such metrics and the inferences made using them. These kinds of metrics are sometimes touted for diagnostic purposes, for example to identify students with relatively low gains who may need further support (Betebenner et al., 2016). However, even for this laudable purpose, some allowance for measurement error ought to be made or many false positives could arise.

The exposition so far on relative gain or conditional status metrics for individual students may seem disparaging. Nonetheless, these metrics can be helpful for understanding the range of gain scores that are historically ‘typical’ for students with similar prior performance in the given measurement context. The following conditions also go some way towards increasing the reliability of conclusions based on these metrics and might make reasonable preconditions for their adoption in practice:

- ensuring assessments are well-targeted for all individual students, for instance through adaptive assessment designs
- incorporating additional prior scores when constructing ‘like groups’ against which to compare gains
- triangulating other evidence about learning progress in the same domain.

These metrics are much less impacted by measurement error and therefore more reliable when aggregated across many students. However, even when aggregated, they are not completely free of bias and care should be taken in their analysis (Lockwood & Castellano, 2017).

So, what is there to gain?

Gain information is more readily available than robust growth trajectory information, but it is inherently volatile and subject to biases that complicate its use. These limitations beg the question of just how much weight to give to individual student gain metrics in practice, whether absolute or relative, for monitoring and responding to evidence about an individual student’s progress.

Viewing the two consecutive scores as two of many along a longer-term progression of increasing knowledge and skill provides more solid footing. This is consistent with the growth mindset advocated by Masters (2016). This frame of reference could include described proficiency levels or learning progression levels or qualitative achievement standards. Given that each level or band occupies a scale score interval usually much larger than a standard error of measurement, these criterion-referenced or standards-referenced progressions provide much more stable markers of progress.

The availability of gain information from numerous assessments invites critical reflection. In the absence of more robust growth trajectory information, absolute gain measures and their normative derivatives might best be incorporated with caveats to augment substantive interpretations of individual student progress. Without this additional score information or this stable, longer-term frame of reference for learning progress, it seems there is little more to gain.

References

Australian Curriculum, Assessment and Reporting Authority. (2019). NAPLAN achievement in reading, writing, language conventions and numeracy: National report for 2019, ACARA.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Betebenner, D.W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Betebenner, D. W. (2011, 19 June). *New directions in student growth: The Colorado growth model*. Paper presented at the National Conference on Student Assessment, http://ccsso.confex.com/ccsso/2011/webprogram/Session2199.html

Betebenner, D. W. & Linn, R. L. (2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda.

Castellano, K. E., & Ho, A. D. (2013a). A Practitioner's guide to growth models. Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and Quantile Regression Approaches to Student 'Growth' Percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 190–215. https://doi.org/10.3102/1076998611435413

Cronbach, L. J., & Furby, L. (1970). How we should measure change—or should we? *Psychological Bulletin*, *74*, 68–80.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently asked questions about growth curve modeling. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Department of Education and Training. (2018). *Through growth to achievement: Report of the Review to Achieve Educational Excellence in Australian Schools*. Commonwealth of Australia. https://www.dese.gov.au/quality-schools-package/resources/through-growth-achievement-report-review-achieve-educational-excellence-australian-schools

Hollingsworth, H., Heard, J., & Weldon, P. R. (2019). *Communicating Student Learning Progress: A Review of Student Reporting in Australia*. Australian Council for Educational Research. https://research.acer.edu.au/ar_misc/34

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kuczmarski, R. J., Ogden, C. L., Guo, S. S., Grummer-Strawn, L. M., Flegal, K. M., Mei, Z., Wei, R., Curtin, L. R., Roche, A. F., & Johnson, C. L. (2002, May). *2000 CDC growth charts for the United States*. National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_11/sr11_246.pdf

Li-Grining, C. P., Votruba-Drzal, E., Maldonado-Carreño, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology*, *46*(5), 1062–1077. https://doi.org/10.1037/a0020066

Lockwood, J. R., & Castellano, K. E. (2017). Estimating true student growth percentile distributions using latent regression multidimensional IRT models. *Educational and Psychological Measurement*, *77*(6), 917–944. https://doi.org/10.1177/0013164416659686

Masters, G. N. (2016, 25 July). Monitoring student growth. *Teacher.* https://www.teachermagazine.com.au/columnists/geoff-masters/monitoring-student-growth

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, *34*(1), 15–21.

Merton, R. K. (1968, 5 January). The Matthew Effect in science. *Science*, 56–63.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306–321. https://doi.org/10.1177/0022219408331037

Nese, J. F. T., Lai, C-F., & Anderson, D. (2013). *A primer on longitudinal data analysis in education. Technical report #1320*. Behavioural Research and Teaching. University of Oregon. https://files.eric.ed.gov/fulltext/ED545257.pdf

O'Malley, K., Murphy, S., McClarty, K., Murphy, D., & McBride, Y. (2011, September). *Overview of student growth models*. White Paper. Pearson.

Patz, R. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Council of Chief State School Officers.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sireci, S. G., Wells, C. S., & Keller, L. A. (2016). *Why we should abandon student growth percentiles.* (Research Brief No. 16-1). Center for Educational Assessment. http://www.umass.edu/remp/news_SGPsResearchBrief.html

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew Effects in reading. *Review of Educational Research*, *84*(2):203–244. https://doi.org/10.3102/0034654313509492

Ployhart, R. E., & MacKenzie, W. I., Jr. (2015). Two waves of measurement do not a longitudinal study make. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 85–99). Routledge/Taylor & Francis Group.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Willett, J. B. (1994). Measuring change more effectively by modeling individual change over time. In T. Husen, & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed.). Pergamon Press.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*(1), 59–69. https://doi.org/10.1177/014662169602000106

Williamson, G. L. (2018). Exploring reading and mathematics growth through psychometric innovations applied to longitudinal data, *Cogent Education*, *5*(1), 1464424, https://doi.org/10.1080/2331186X.2018.1464424