A large-scale, long-term study of scale drift: The micro view and the macro view

W He,[1] S Li,[1] G G Kingsbury[2]
[1]NWEA, 121 NW Everett St, Portland, OR 97209
[2]Independent consultant, 6134 SE Lincoln, Portland, OR 97215

E-mail: wei.he@nwea.org

Abstract. The development of measurement scales for use across years and grades in educational settings provides unique challenges, as instructional approaches, instructional materials, and content standards all change periodically. This study examined the measurement stability of a set of Rasch measurement scales that have been in place for almost 40 years. In order to investigate the stability of these scales, item responses were collected from a large set of students who took operational adaptive tests using items calibrated to the measurement scales. For the four scales that were examined, item samples ranged from 2183 to 7923 items. Each item was administered to at least 500 students in each grade level, resulting in approximately 3000 responses per item. Stability was examined at the micro level analysing change in item parameter estimates that have occurred since the items were first calibrated. It was also examined at the macro level, involving groups of items and overall test scores for students. Results indicated that individual items had changes in their parameter estimates, which require further analysis and possible recalibration. At the same time, the results at the total score level indicate substantial stability in the measurement scales over the span of their use.

1. Introduction

Developing and maintaining vertical scales for the measurement of educational achievement is a complex task. Beyond the psychometric issues, the educational setting brings its own set of challenges to accurate measurement. Among these are the following:

- From grade to grade, the content that a student is learning may change noticably, which will change the questions that it is appropriate to administer.
- From school to school, the amount and order of content changes noticably, even for students in the same grade.
- From year to year, schools may change their curriculum, instructional approaches, and instructional materials.
- As they grow, students change in their willingness to engage in assessments, and their interest in the subject being measured.
- As time passes, federal and state regulations change the requirements placed on the schools and their students.

The nature of education is an ecosystem that is constantly changing, so creating a measurement scale that measures student achievement in a consistent fashion is an interesting challenge.

In the late 1970s, the Northwest Evaluation association (NWEA) began to develop a set of measurement scales to examine student achievement in mathematics and reading. The initial work to develop these scales attached items to a Rasch measurement scale using the four-square linking design

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

developed by George Ingebo [1] and explicated by Ben Wright [2]. The purpose of the initial development was to create item banks that could be used by multiple school districts to create assessments that matched their curriculum. The original items were associated with content from grades 3 to 8, and all assessments were administered in paper form to students in the pacific northwest.

Since that time, NWEA has expanded and updated the item banks, extended the range of the scales to measure students from kindergarten to high school, added new item styles, added a host of accommodations, moved to adaptive testing, moved to electronic delivery of all tests, and expanded to work with students across the United States, Canada, and a number of schools in other countries. Given these changes, it makes sense to examine the stability of the measurement scales, to assure that a student with a particular level of achievement would obtain approximately the same score today that they would have obtained several decades ago. A previous study by Kingsbury and Wise [3] indicated that the NWEA measurement scales in reading and mathematics showed little drift in their first twenty years of use. The current study expands on that earlier study by adding fifteen years of data and adding the language usage scale and the general science scale.

2. Methodology

A total number of 26,059 items were used in this study. These items were originally calibrated between January, 1980 and January, 2014. The items were recalibrated using responses gathered during January 1st, 2014 and November 30th, 2015. To perform the recalibration, at least 1,000 responses were collected for each item in an operational test. On average, items in each item pool contained an average of 2,953 student responses (with at least 500 responses in each grade). Since the operational tests were adaptive, items differed in their frequency of use and also in the achievement levels of students to whom they were administered. Table 1 presents the sample size of the item pool for each subject as well as the average number of total student responses per item used for the recalibration.

Table 1. Description of Item Pools and Student Responses

Scale	Items	Average number of responses per item
Reading	7776	2949
Math	10106	2941
Language Usage	5357	2956
General Science	2820	2967

2.1. Analysis: The micro view

One approach to examining scale drift is to identify items associated with the scale that change their difficulties over time. In an educational setting, this type of change might occur for any number of reasons, including change in instructional emphasis, change in common usage of words, or change in the underlying scale. This examination of item parameter stability serves as the micro view of scale stability. Consistent item calibrations allow us to conclude that the scale is remaining stable, but changes in item parameter estimates may spring from a variety of sources.

This study used the robust Z statistic developed by Huynh & Rawls [4] that originates from robust statistical procedure to detect unstable items. Z-scores are expressed in terms of standard deviations from the mean. As a result, z-scores have a distribution with a mean of 0 and a standard deviation of 1. Z score is calculated by

$$z = \frac{X - \mu}{\sigma}$$

where X is an individual score, μ is the mean, and σ is the standard deviation.

In order to compute the robust z, the mean is replaced by the median and the standard deviation is replaced by 0.74 times the interquartile range (IQR). The quantity 0.74*IQR is used to match the standard deviation of the normal distribution.

The robust statistic for each item is the ratio $z = (D - Md) / (0.74 \times IQR)$, where D is the difference between original b-parameter and the newly calibrated b-parameter, Md is the median, and IQR is the interquartile range for the differences. α in each direction was set at 5%, and the critical value is z*=±1.645, correspondingly. All items with a robust z smaller than the absolute value of z* in absolute value were regarded as stable. Otherwise, items were flagged as drifting. This approach should identify approximately 10% of items as drifting if the null hypothesis is true. This allows the identification of many items for review, ensuring that any item with noticeable drift can be examined by content experts.

2.2. Analysis: The macro view

An additional approach to examining drift is to look at the impact of parameter estimate changes on the actual outcome measures, the total test scores for the students. This macro view accepts that individual items may drift, but asks whether this drift has an impact on students' scores and the educational decisions made with them.

Since the tests used in this analysis were adaptive, the impact of drift on total test scores will vary from student to student. In order to capture the overall impact and the variance of that impact, a set of 1,000 20-item adaptive test events were simulated using the original calibrations to select items from the item pool for General Science. General Science was selected for the macro analysis since it had the greatest percentage of items flagged for drift during the micro analysis (see below). Simulated responses were based on the new calibrations. Each test event was then scored twice, once using the original parameter estimates, and one using the new parameter estimates. These two sets of scores were then compared to identify how the change in calibration estimates would impact student scores and subsequent decisions in the classroom.

3. Results

3.1. Micro results

The existing item difficulty parameters in the bank with those from the new calibration were correlated with each other. As Table 2 indicates, the correlation coefficients $r_{old,\ new}$ ranged from .97 and .99 for the different scales. Using the Robust Z method (α=10%), 13-24% of the items were flagged for parameter drift for the different scales. These percentages are larger than the 10% expected under the null hypothesis, and they indicate that some amount of item drift is occurring in the items associated with each measurement scale.

While Table 2 indicates that a greater than expected proportion of the items are flagged for drift, it also indicates that the correlations between the original and new calibrations are extremely high for each of the measurement scales. This suggests that the ordering of the item difficulties is quite similar to that identified by the original calibrations.

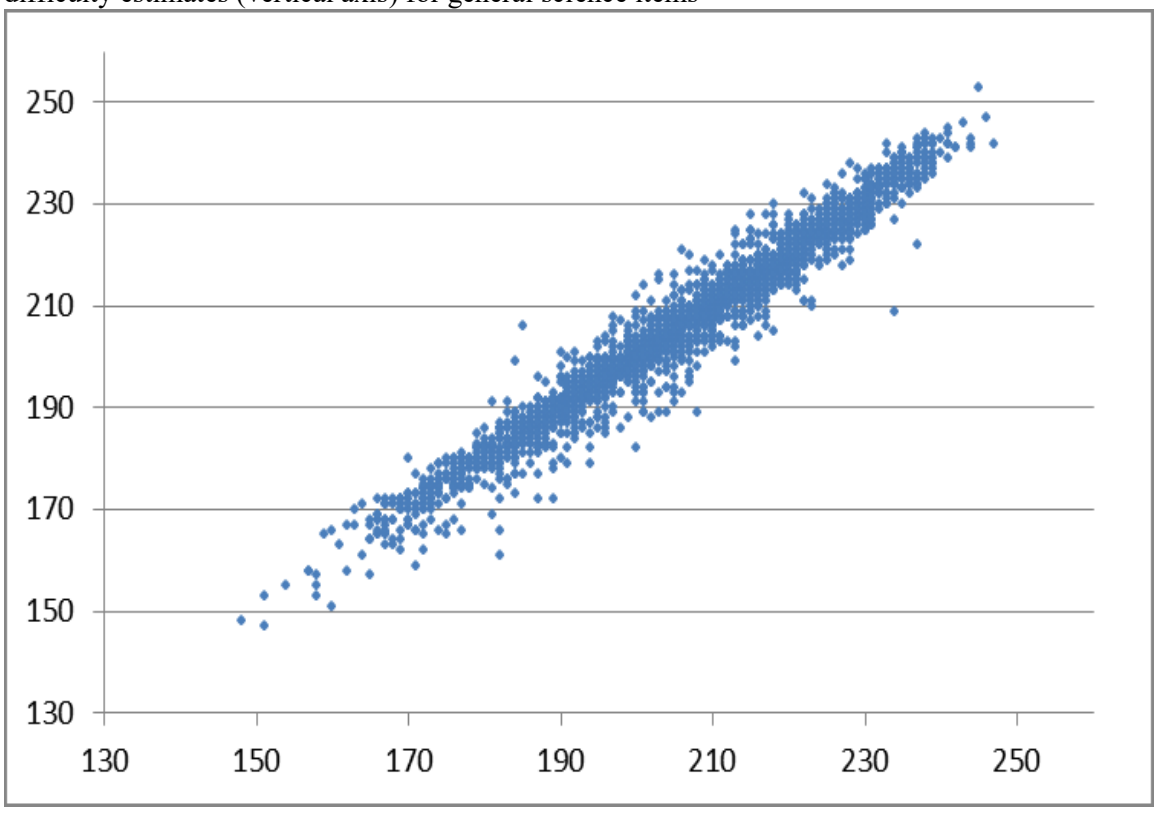
Table 2. Item Parameter Drift Detection

Item Pool	**N of items**	**N_flagged**	**% flagged**	$r_{old,\ new}$
Reading	7776	1191	15%	0.98
Math	10106	1958	19%	0.99
Language Usage	5357	693	13%	0.97
General Science	2820	681	24%	0.97

On average, approximately 19% of mathematics items flagged for drift. Reading items tended to be more stable than math and general science items. On average, approximately 15% of reading items were flagged as drifting, while approximately 19% of math items and 24% of general science items were flagged respectively. Since the general science items show the highest percentage of items flagged as drifting, this scale will be used as the example for further analysis.

Figure 1 shows the original calibrations for each of the general science items plotted against the new calibrations for the same items. It can be seen that the vast majority of the items have new calibrations that are quite similar to their original calibrations. Over 2000 points are on the graph, most of which have new and original calibrations that differ by two or fewer scale score points (or less than .2 logits). It is also clear that a small number of items have a new calibration that differs substantially from the original calibration (several by 2.0 logits or more). If the authors of the scale do not wish to recalibrate all of the items at once (which might lead to disruptions of trend information) they can use this figure to triage the drift to allow sequential review and recalibration.

Figure 1: Original item difficulty estimates (horizontal axis) and new item difficulty estimates (vertical axis) for general science items



3.2 Macro Results

Since the Micro analysis indicated that the scale associated with general science had the highest percentage of items indicating parameter drift, this scale was chosen for the Macro analysis. For this scale, 24% of the items were flagged as drifting, so it is useful to ask how the drifting items influence overall test scores for students.

The simulation of 1000 adaptive test events indicated the following:

- The average difference between the scores based on the original calibrations and the scores based on the new calibrations was -0.011 logits.

- The average absolute difference between the scores based on original and new calibrations was 0.061 logits.
- The lowest difference between the scores based on original and new calibrations was -0.28 logits.
- The greatest difference between the scores based on original and new calibrations was 0.26 logits.

It is useful to note that in this simulated sample, the standard deviation of general science test scores was 1.73 logits.

4. Conclusions

The use of the micro and macro analyses allow us to look at the stability of measurement scales from different perspectives. The micro analysis allows us to look at individual items and sets of items, to determine whether they continue to measure the construct of interest in the same manner over a period of time. This allows us to identify whether processes for mid-course corrections need to be put into place for specific items in order to maintain accurate measurement. This is the part of the measurement words in which psychometricians tend to spend most of their time, and decisions here are crucial for the ongoing health of a measurement scale. In this study, the micro analysis indicates that up to 24% items should be further reviewed, to determine whether calibration values should be adjusted.

The macro analysis allows us to observe the impact that changes at the micro level have on decisions made using total test scores with groups of students. Since students see a sampling of items in any adaptive test, it is likely that they will see mostly items that haven't drifted, as well as a few items that have drifted, becoming more difficult or less difficult. It is clear from the simulation that the impact of the drifting items is quite small. The greatest differences that were observed for total test scores were less than .3 logits. This is equivalent to 3 points on the reported score scale, or less than a fifth of a standard deviation in student achievement. The average difference in scores was less than one point on the reported score scale (smaller than the smallest difference reported).

Two of the common uses of scores from this test are instructional grouping and identification of instructional need. Each of these score uses involves dividing a class into two to five instructional groupings. While these groupings are critical to good instruction, it is very unlikely that any students would be misplaced due to item drift.

This study provides strong evidence that these vertical scales in education need periodic review to identify items that may be drifting in difficulty, and need a process to deal with items that are identified. It also provides strong evidence that the drift that has occurred in the general science scale (the highest level of drift observed) does not have a substantial impact on the total test scores for students or on the decisions made from these scores.

5. References

[1] Ingebo G 1997 *Probability in the measure of achievement* (Chicago, IL: MESA Press)
[2] Wright B 1977 Solving measurement problems with the Rasch model *Journal of Educational Measurement* **14** 97-116
[3] Kingsbury G and Wise S 2011 Creating a K-12 Adaptive Test: Examining the stability of item parameter estimates and measurement scales. *Journal of Applied Testing Technology*
[4] Huynh H and Rawls A 2009 A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting* ed E V Smith Jr and G E Stone (Maple Grove, MN: JAM Press)