dr Joanna Tomkowicz
CTB/ McGraw-Hill, Kanada

Konstrukcja i zastosowanie skal pionowych w pomiarze osiągnięć uczniów

Abstrakt przygotował: B. Niemierko

Skale pionowe (*vertical scales*) obejmują kontinuum rozwojowe osiągnięć uczniów w kolejnych etapach uczenia się, a w szczególności – w kolejnych klasach szkolnych. Rosnącej trudności testy są w nich *wiązane* (*linked*), ale nie mogą być *zrównywane* (*equated*), jak *poziome* (*horizontal*) wersje równoległe, gdyż mierzone umiejętności ulegają w nich systematycznym zmianom. Te zmiany są umiarkowane w zakresie czytania i matematyki, ale w zakresie przyrody i nauk społecznych są tak duże, że treść serii testów traci ciągłość.

Artykuł przedstawia założenia i procedury konstrukcyjne skal pionowych z języka ojczystego i matematyki przeznaczonych do zastosowania w programie „Żadne dziecko nie zostaje w tyle" (*No Child Left Behind*). Ten program, ustawowo wprowadzony w Stanach Zjednoczonych w 2001 roku, zobowiązuje szkoły do zapewnienia wszystkim uczniom *pełnych umiejętności* (*proficient level*) w zakresie tych dwu przedmiotów. Ustawa wyznacza termin 2013/2014 na osiągnięcie celu i nakazuje zbliżanie się do niego rocznymi etapami udokumentowanymi obowiązkowym pomiarem testowym w klasach od III do VIII.

Istnieją trzy wzory procesu konstrukcyjnego skali pionowej osiągnięć uczniów:

1. *Zrównywanie skal*. Zrównanie nowej skali z istniejącą już skalą pionową wytworzoną przez niezależnie działające duże przedsiębiorstwo pomiarowe.
2. *Wspólne zadania*. Zastosowanie wybranych *zadań kotwiczących* (*anchor items*), jednakowych w co najmniej dwu kolejnych klasach. Ten wzór był zastosowany w przedstawianym tu programie, a uczniowie z klas IV-VII rozwiązywali po dziesięć zadań wyboru wielokrotnego z klasy niższej i z klasy wyższej, zaś uczniowie klasy III – piętnaście zadań z klasy IV i uczniowie z klasy VIII piętnaście zadań z klasy VIII, dwa tygodnie po właściwym testowaniu.
3. *Wspólni egzaminowani*. Zastosowanie, obok testu właściwego dla danej klasy, testu przeznaczonego dla klasy programowo niższej lub wyższej, bez wliczania dodatkowych punktów do wyniku ucznia. Ten wzór był zastosowany w przedstawianym tu programie, a uczniowie z klas IV-VII rozwiązywali po dziesięć zadań wyboru wielokrotnego z klasy niższej i z klasy wyższej, zaś uczniowie klasy III – piętnaście zadań z klasy IV i uczniowie z klasy VIII piętnaście zadań z klasy VIII, dwa tygodnie po właściwym testowaniu.

Wśród procedur budowana skali pionowej wyróżniają się metody oparte na *teorii wyniku zadania* (*Item Response Theory*), polegające na *kalibrowaniu zadań* (*item calibration*) w drodze szacowania ich parametrów (*a*) mocy różnicującej, (*b*) trudności, i – w przypadku zadań zamkniętych – (*c*) poziomu zgadywania. Tu mamy do wyboru dwie metody:

1. *Kalibrowanie osobne i wiązanie łańcuchowe* (*chain linking*). Obejmuje to wybór *klasy bazowej* (*base grade*), najlepiej w środku szeregu – w przedstawianym programie była to klasa V – i, rozpoczynając od tej klasy, wiązanie klas wyższych i klas niższych przez wspólne zadania. W końcu wyniki są przetwarzane ze skali *teta* (średnia 0, odchylenie standardowe 1) na dowolną skalę standardową – w tym wypadku na skalę o średniej 500 i odchyleniu standardowym 40. Ta metoda została zastosowana w podanym przykładzie.
2. *Kalibrowanie jednoczesne* (*concurrent calibration*), w jednej puli zadań i z jednym zbiorem parametrów. Ta metoda została zastosowana w podanym przykładzie.

Wyniki uczniów mogą być przedstawiane także na dwa sposoby:

1. W wybranej *skali punktowej*, przy założeniu monotonicznego wzrostu *normy* ilościowej (średniej arytmetycznej), ale niekoniecznie równych przedziałach kolejnych klas.
2. W układzie *poziomów wymagań* (*proficiency levels*), zwykle w czterech kategoriach (*not proficient – partially proficient – proficient – advanced*), a więc zgodnie z założeniami *pomiaru sprawdzającego* (*criterion-referenced measurement*). Tu normy ilościowe ustala zespół ekspertów przedmiotowych, biorąc pod uwagę zarówno (stanowe) standardy edukacyjne, jak i rozkłady wyników punktowych próbnych zastosowań skali pionowej.

W podanym przykładzie zastosowano do wyników zadań *analizę czynnikową* metodą głównych składowych, która potwierdziła zakładaną jednowymiarowość przestrzeni zmiennych, przy czynniku *zdolność* (*ability*) jako czynniku głównym. Przeprowadzono klasyczną *analizę zadań*, uzyskując wskaźniki łatwości w granicach 0,20 – 0,95 i moc różnicującą (współczynnik korelacji punktowo-dwuseryjnej) powyżej 0,15 dla wszystkich zakwalifikowanych zadań. Następnie dokonano przekształcenia średnich i odchyleń standardowych zadań przeznaczonych dla poszczególnych klas szkolnych, tak by uzyskać średnią (500) i odchylenie standardowe (40), przewidziane dla klasy bazowej skali pionowej (Tab. 1.).

Krzywe charakterystyczne sześciu zbudowanych testów z języka ojczystego i matematyki oraz *błędy standardowe pomiaru* w różnych częściach skal są przedstawione na Rysunkach 1. i 2. Średnie i odchylenia standardowe wyników w skali pionowej przedstawiają Tabele 2. i 3. oraz Rysunki 3. i 4. Nadto w Tabelach 2. i 3. podano wartości *minimalne* i *maksymalne* możliwe do uzyskania w poszczególnych klasach. Wartości wybranych *centyli* danych empirycznych są zawarte w Tabelach 4. i 5., a *krzywe kumulacyjne* rozkładów liczebności pokazują Rysunki 5. i 6.

Przeglądając te dane, zauważamy:

1. bardzo niską wartość *efektu standardowego* (*effect size*) klasy szkolnej: średnio tylko 0,34 odchylenia standardowego dla języka i 0,39 dla matematyki, co wynika zapewne z ilościowej przewagi zadań zamkniętych (wyboru wielokrotnego),
2. trend *krzywoliniowy*, kurczące się przedziały skali (silniej dla języka ojczystego, słabiej dla matematyki) i *ujemny efekt wachlarzowy* (spadek rozrzutu) – to wszystko spowodowane zapewne *efektem pułapu* (*ceiling effect*), wynikającym z charakteru elementarnego mierzonych umiejętności.]

Autorka kończy artykuł następującymi przestrogami:

1. *Błędy pomiaru* osiągnięć uczniów skalą pionową są stosunkowo duże (Rys. 1. i 2.), zwłaszcza na jej krańcach.
2. Rozróżnienia są zadowalająco *rzetelne* raczej dla grup niż dla pojedynczych uczniów, bo różnica osiągnięć ucznia uzyskana w ciągu jednego roku szkolnego może okazać się w większości przypadków nieistotna statystycznie.
3. Różnice między kolejnymi pomiarami osiągnięć ucznia lub grupy uczniów mogą być wyolbrzymione przez zjawisko *regresji ku średniej*.
4. Skala starzeje się i jej *utrzymanie* (*maintenance*) wymaga systematycznej wymiany części zadań.
5. Zastosowanie podobnego podejścia w Polsce wymagałoby wiązania kolejnych klas, a nie szczebli szkoły (6 – 9 – 12 lat kształcenia), gdyż zmiany treści odpowiednich umiejętności (ewolucje odpowiednich *konstruktów*) między szczeblami szkoły wydają się zbyt duże.
6. Budując taką skalę, powinniśmy skupić uwagę na zgodności treściowej – przy odpowiednio zróżnicowanej trudności – zadań przeznaczonych dla kolejnych klas, a więc na *trafności wewnętrznej* pomiaru określonej umiejętności.

Development and Use of Vertical Scales to Measure Student Achievement

Introduction

Standardized educational achievement tests are used to provide valid, reliable, and objective information about student achievement in different areas of learning. When a test is administered, individual or group results may be compared with a criterion or with the results of other students in the class, in the school, in the districts, or in a notional norm group. Given an appropriate test design, the test results can also be used to make comparisons over time and measure student progress as they move on a learning continuum. This kind of comparison can provide one indication of academic growth for individuals or groups of students. When the test results are used in conjunction with other measures, such as classroom observation, teacher-made tests, external assessments can provide valuable information about the progress of students, as well as the effectiveness of educational programs.

In the United States, every state is federally mandated to test in grades three through eight in Reading and Mathematics. The No Child Left Behind Act of 2001 (NCLB; Public Law 107-110) requires all states to have Reading and Mathematics testing programs implemented by the 2005-2006 academic year. This Act specifies goals for adequate yearly progress (AYP) and calls for all students reaching at least the proficient level by 2013-2014 academic year.

Prior to the implementation of NCLB, most states tested at only a few grades and only a few states tested contiguously across grades. When several years separate testing grades, monitoring student yearly progress based on the test results is difficult. However, when tests are administered in continuous grades, an appropriate test design called vertical scaling permits educators to make reliable and accurate inferences about student achievement across grades and over time.

Vertical scale concept

Vertical scaling of assessments administered in continuous grade level provides a tool to measure student capabilities and skills as they move up the grade level. A vertical (or developmental) scale, which can be viewed as a developmental continuum, is often seen as helpful in assessing the growth of students from one grade level to the next as demonstrated by their scale scores. These scale scores on the vertical scale, when properly constructed and maintained, represent units on a single, equal-interval scale applied across all grade levels. If a test (i.e., Mathematics) is placed on a vertical scale, test scores can be directly compared from year to year, albeit with some constraints. Student growth can be measured

simply by subtracting the previous grade's scale score from the current grade's scale score. The scale score difference in achievement from one year to another indicates the amount of student growth. This basic concept of a vertical scale is fairly straightforward.

Vertical scaling can also be conceptualized as a measurement process that models latent variable estimates derived from a set of test forms of increasing difficulty, as well as puts the ability estimates in appropriate relation so that comparisons may be made for examinees taking forms of different difficulty. Creating vertical scales involves linking test forms to a common scale. Note that test linking here is not equivalent to test equating. Formally equating forms, which results in interchangeable test scores, requires that forms be parallel in their content and technical characteristics. Forms that are not parallel in structure but measure a common proficiency may be linked (Patz, 2007.) Overlap in content standards at adjacent grades may support the assumption that forms for adjacent grades measure a common construct. Differences in the standards and psychometric properties of the test forms (e.g., test difficulty) mean that these forms are not parallel and so they may be linked. Since the links used in this case relate forms of intentionally different difficulties, they are referred to as vertical links, and the resulting scale is called a vertical scale. This is in contrast to the linking of test forms of equivalent difficulty, which can be called horizontal linking or equating.

There are several psychometric designs that can be utilized to build a vertical scale and all of them rely on appropriate test framework, namely, continuous test content and learning objective across grades. It is crucial for vertically scaled test interpretation that the changes in the construct being measured across grades be considered when developing a vertical scale (Huynh and Schneider, 2004.) Typically, the degree of construct change from grade to grade differs, depending on a subject area. A vertical scale captures a common dimension across grade levels; it does not capture grade-specific dimensions that may be of importance (i.e., Lissitz and Huynh, 2003.) It has been demonstrated that the greatest changes across grades in the measured construct occur in such subject areas as Social Studies or Science, and for such assessments vertical scales are more difficult to justify. Reading/Language Arts and Mathematics are typically considered continuous, with overlapping constructs at least at each adjacent grade level. As such, it is often perceived that vertical scales for these two subject areas can provide more reliable examinee scores.

Vertical scale development

This section of the paper presents selected procedures for data collection, and scaling and equating methodologies commonly used in vertical scale development.

Data collection designs

Different data collection designs can be employed to gather data for vertical scale development. Three basic designs are presented and discussed in this section. One of them utilizes existing external vertical scale (when available) and two involve building a vertical scale from the scratch.

Equating to existing vertical scale. In this design, an existing vertical scale (i.e., CTB/McGraw-Hill's *TerraNova* Reading scale) is available to use. Such a scale is typically built by a large testing company that has collected student data containing responses to items that are written to specific content requirements. These content requirements come from an examination of many states' curricula. The resulting scale is standardized and norms are developed. A new customized vertical scale can be developed by embedding items from the existing vertical scale into a new custom test. The embedded items may or may not contribute to the final student score on the new test but will be used as anchor items to equate the new tests to the existing vertical scales in a process of horizontal on-grade level equating. These anchor items must conform to the new test content specifications. The new tests equated to the existing vertical scale will also be on a vertical scale. Because the new tests are equated to the existing vertical scale, the between-grade links for these tests are assumed based on this design. A final linear transformation of the new tests on vertical scale can be performed to differentiate the new scale from the existing scale in terms of their psychometric properties.

Common item design. In this design a new scale is built using common items shared by adjacent forms to link adjacent grade levels. The number of common items between adjacent levels, which may be called "vertical anchors", should be sufficiently large and the items should be representative of the domain of adjacent grade test contents to provide a reliable link. Under this design the vertical anchors are embedded in tests on which students will be scored and the anchors are selected based on the content standards shared between grade levels. The vertical anchors can be selected either from both below and above grade levels or only from the grade level below. If vertical anchors are well aligned with the adjacent grade level content specifications, it does not matter whether the item was originally intended or is operationally being used at grades above or below. Once the vertical anchors are administered along with the regular test items to students, they provide link between adjacent grade levels. The vertical anchors may or may not contribute to the total test score.

Common examinee design. This design is typically employed when the vertically-scaled tests have no items in common. Under this design, students may take two forms of the test—one form targeted for their grade level and another form targeted for an adjacent level. This "common examinee" approach has been used in the creation of nationally standardized, vertically-scaled achievement test batteries (e.g., CTB/McGraw-Hill's *TerraNova, CTBS*, and *CAT* batteries). In those settings, students might take a test level above their grade if the vertical scaling study is conducted in the spring, or a test level below their grade level if the study occurs in the fall (Patz, 2007.) The entire test forms administered to adjacent grades would serve as vertical anchors in linking those grades, but students would not be scored on those forms.

Most state testing programs either utilize an existing vertical scale to build their own custom scale or develop the scale under the common item design. The first option is very straightforward and does not require administration of the off-grade level items to create the between-grade links, but the state custom tests must be similar in their content domain coverage to the existing vertically scaled tests to be able to select appropriate set of anchor items for equating. The second and third options are more complex designs and require administration of items or entire forms from adjacent grade levels in addition to on-grade level forms, but the new scale is built from the scratch and the content coverage of its tests does not need to be aligned with any existing test batteries.

Scaling procedures

A variety of scaling procedures can be used to create vertical scales. This section focuses on Item Response Theory (IRT) scaling procedures. Item Response Theory is a statistical methodology that takes into account the fact that not all test items are alike and that all items do not provide the same amount of information in determining how much a student knows or can do. Computer programs that implement IRT models use actual student data to estimate the characteristics of the items on a test, called "parameters." The parameter estimation process is called "item calibration." Using IRT methodology, one-, two-, or three-parameter logistic models (Lord, 1980) could be used to estimate parameters for the multiple-choice items. The one- or two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) could be used to estimate parameters for the constructed response items. The one-parameter model estimates the item difficulty parameter only; the two-parameter model estimates the item discrimination and difficulty parameters; and the three-parameter model estimates the item discrimination, difficulty, and, for multiple-choice items, the 'guessing' parameters. The discrimination parameter is an index of how well an item differentiates between high-performing and low performing students. An item that cannot be answered correctly by low-performing students, but can be answered correctly by high-performing students, will have a high-discrimination value. The difficulty parameter is an index of how easy or difficult an item is. The higher the difficulty parameter is, the harder

the item. The guessing parameter is the probability that a student with very low ability will answer the item correctly. Once the item parameters are estimated, the equating of the new form to the existing form or between-form linking is performed. Again, a variety of methods including mean/mean, mean/sigma or test characteristic curve (TCC) method (Kolen & Brennan, 2005) could be used for this purpose.

In a development of a new vertical scale any of these methods can be used in its appropriate context. If an existing vertical scale is utilized to build a new scale, a strict horizontal on-grade level scaling and equating is conducted. In this approach, each grade level test is scaled separately and equated to the corresponding grade level test on a vertical scale via anchor items selected from that test and embedded in new test.

If a new scale is being created, the most frequently considered options for item calibration and scaling are a) separate calibrations and chain linking, and b) concurrent calibration. These procedures are briefly described below.

Separate calibrations and chain linking. The separate calibration and chain linking method is accomplished in two steps. The first step is separate calibration of each grade starting from the arbitrary selected base grade (preferably in the middle of continuous grade levels). The second step is grade-by-grade chain linking using common items between grades, again starting from the base grade. The separate calibrations result in the establishment of a unique theta metric scale for each grade. The common items used for linking adjacent grades allow for the development of a common scale. The item parameter estimates for common items are used to estimate scale transformation constants that in turn allow placement of item parameters from each adjacent grade onto the base grade scale using an equating technique. This step is then repeated for each adjacent grade until all grades are placed on the common scale. The initial vertical scale is typically developed in a theta metric (with a mean of 0 and standard deviation of 1) and later transformed to a scale score metric with an arbitrary set mean and standard deviation for base grade.

Concurrent calibration. Concurrent calibration is a method that allows for establishing the common scale in a single step—the calibration phase—by simultaneously estimating parameters for all items at all grades. The estimated parameters in the theta metric are on the same scale. In addition, population ability estimates are obtained for multiple groups (separate grade level). The group mean and standard deviation for the base grade are then used to derive transformation constants and to transform the initial parameter estimates in the theta metric into the common scale score metric. No cross-grade equating is necessary and the same transformation parameter constants are applied to item parameter estimates across all grades.

Student scores

A variety of derived student scores can be obtained from vertically scaled assessments including criterion referenced scores and normative scores. The criterion referenced scores include scale score and performance level scores. In the United States, these scores are used in state assessments to evaluate student achievement against established criterion, and to measure year-to-year progress against the same. The norm referenced scores (i.e., percentiles, stanines, normal curve equivalents, and grade equivalents) are typically not used in custom state assessments and are not discussed in this paper.

Scale scores

Once the vertical scale is established the item parameter estimates are used to derive student scores. Typically, only on-grade level items contribute to a test score, but occasionally, the common between-grade items are also used in individual student score estimation. Different scoring methods can be used to derive student scores. For example, the raw score on the test can be converted to a scale sore by means of a conversion table or the item pattern method can be employed. However, regardless of item source (on-grade level only or combination of on-grade and off-grade level) or the scoring method used to derive student scores, there is an expectation of grade-to-grade growth as expressed by increasing group (or population) scale score mean. In other words, the mean scale scores are expected to increase by a number of scale score units from grade to grade, indicating student progress on learning continuum. There is no expectation of linear growth. In fact, most of the vertical scales indicate non-linear growth with more academic growth observed in lower grade levels and less growth observed in higher grade levels.

Proficiency levels

Following the vertical linking of consecutive grade level assessments the cut scores defining student proficiency levels can be established. Such cut scores are test criteria to which individual student scores are related and are typically established by state educators during the standard setting workshop. Most often, three cut scores are set for each grade and they divide the scale into four proficiency levels: not proficient (not meeting learning standards); partially proficient (partially meeting learning standards); proficient (meeting learning standards); and advanced (exceeding learning standards). The cut scores are set based on test content and educators' judgment supported by the empirical data of what a partially proficient, proficient, and advanced student should know and be able to do. It is expected that the cut scores on vertically scaled assessments will be vertically moderated or, in other words, logically progressing across grade levels resulting in a smooth and rational pattern of percent of students falling into each proficiency category. There are two primary conditions that must be met to establish vertically moderated standards (VMS). First, a set of common policy definitions for the achievement levels needs to be used for all grades.

Second, a consistent trend line needs to be imposed on the percentage of students in proficiency levels across grades. In the VMS approach student growth could then be measured from year to year by measuring a student's progress relative to proficiency. In other words, a student's yearly progress is defined in terms of adequate end of year performance that allows the student to successfully meet the challenges in the next grade (Lissitz and Huynh, 2003.)

Example of a vertical scale development

This section provides an example of a vertical scale development for a large scale assessment programs in English Language Arts (ELA) and Mathematics grades 3-8.

Data collection design

The ELA and Mathematics scales presented in this example were established using a common item linking design. In this design samples of students were administered sets of operational test items from adjacent grades in addition to their regular on-grade level assessments. These off-grade level items were used for linking adjacent grades but did not contribute to the test score. The selected linking items were aligned with the learning standards assessed in grades 3-8 ELA and Mathematics tests. The common standards measured by grades 3-8 ELA test were: Information and Understanding, Literary Response and Expression, and Critical Analysis and Evaluation. The common learning standards measured by grades 3-8 Mathematics tests were: Number Sense and Operations, Algebra, Geometry, Measurement, and Statistics and Probability. Only multiple-choice items were used for linking purposes. Approximately 10 below and 10 above grade level items were administered to samples of students in grades 4 through 7. Grade 3 students were administered 15 above grade level items (from grade 4 test) and grade 8 students were administered 15 below grade level items (from grade 7 test). The regular on-grade level assessments were administered to all students in grades 3-8 during a regular operational administration window. The off-grade level items were administered to samples of students approximately two weeks after the regular operational tests.

Data analysis

The off-grade level data were merged with operational test data using student ID and other student and school information. The samples of students who took linking items were assessed for their representativeness of the population in terms of school geographic location, school Socio-Economic status, student achievement, and student ethnicity. Some sampling down was performed to better align sample characteristics to the state student population characteristics.

To demonstrate the common factor (ability) underlying student responses to on- and off-grade level ELA and Math test items, a principal component factor analysis was conducted on the polychoric correlation matrix of individual items for each data set consisting of on-grade level items and off-grade level items taken

by the same groups of students. A large first principal component was evident in each analysis indicating that the between grade level data were essentially unidimensional thus satisfying the requirement of common construct being measured by at least adjacent grades.

A classical item analysis was performed on the data used for vertical scale development and item difficulties (p-value), item point-biserial correlations, and omit rates were computed. Overall, students performed better on operational on-grade level items than on

above-grade level items for both ELA and Mathematics. Additionally, with exception of grade 6 ELA, they performed better on below-grade level items as compared to their performance on their regular on-grade level operational test items. All items across both content areas displayed reasonable p-values (ranging form 0.20 to 0.95) and point-biserial correlations (above 0.15). Omit rates were generally less than 2%. Due to space limitation dimensionality and item analysis results are not presented in this paper.

Following data dimensionality assessment and classical analyses, IRT methodology was used to scale the ELA and Mathematics assessments. Two IRT models were used to calibrate the operational and vertical scaling test items. The three-parameter logistic (3PL) model was used to estimate parameters for the multiple-choice items. The two-parameter partial credit (2PPC) model was used to estimate parameters for the constructed response items. Marginal Bayesian Estimation was used to estimate item parameters. In the Bayesian approach, the parameter to be estimated is considered a random variable that follows a certain probability distribution (i.e., prior distribution). By taking into account the prior information for the unknown parameter in the estimation process, Bayesian estimation is expected to improve the accuracy of the estimated value (Gao & Chen, 2005; Mislevy, 1986.) The Stocking and Lord (1983) Test Characteristic Curve (TCC) method was used for equating.

A separate calibration and chain linking method was employed to establish a common scale across grades within each content area. For both content areas, grade 5 was identified as the base grade. The operational test data for grade 5 were calibrated first and item parameters and student mean ability in the theta metric were estimated. These estimates were used to identify transformation constants that would permit linear transformation of item parameter estimates from the theta metric into the scale score metric and thus, produce a scale with the desired mean and standard deviation for the base grade. The mean in scale score metric for the base grade was set to be 500 and the standard deviation was set to be 40, for both ELA and Mathematics.

The following formulas were used to compute transformation constants for the transformation of the base grade item parameter estimates from the theta metric to the scale score metric:

$$M1 = \frac{SD_{ss,5}}{SD_{\theta,5}}$$

and

$$M2 = \overline{X}_5 - (\overline{\theta}_5 * M1)$$

where:
M1 and *M2* are the transformation constants
$SD_{ss,\,5}$ is the desired standard deviation in scale score metric for the base grade
$SD_{\theta,\,5}$ is the estimated standard deviation in the theta metric for the base grade
θ_5 is the estimated population mean in theta metric for the base grade
$\overline{X}_5$ is the desired mean in scale score metric for the base grade

The *M1* and *M2* transformation constants were then applied to the base grade item parameter estimates in the theta metric to transform them into scale score metric using the following formulas:

$$A_{ss} = a_\theta / M1$$
$$B_{ss} = M1 * b_\theta + M2$$
$$F_{ss} = f_\theta / M1$$
$$G_{ss} = g_\theta + (f_\theta / M1) * M2$$
$$C_{ss} = c_\theta$$

where:
A_{ss} is a discrimination parameter in scale score metric for MC items
B_{ss} is a difficulty parameter in scale score metric for MC items
F_{ss} is a discrimination parameter in scale score metric for CR items
G_{ss} is a difficulty for category m_j in scale score metric for CR items
a_θ is a discrimination parameter in the original theta metric for MC items
b_θ is a difficulty parameter in the original theta metric for MC items
f_θ is a discrimination parameter in the original theta metric for CR items
g_θ is a difficulty level for category m_j in the original theta metric for CR items
C_{ss} and c_{ss} is a guessing parameter in the original theta metric

Table 1 presents the initial population mean and standard deviation estimates and the transformation constants used for scale transformation of the base grade (5) for ELA and Mathematics.

Table 1. Transformation constants for ELA and Mathematics base grade

Content Area and Grade	Desired scale properties in scale score metric		Estimated sample ability in theta metric		Transformation constants	
	Mean	SD	Mean	SD	M1	M2
ELA 5	500	40	0.05	1.227	32.5998	498.3700
Math 5	500	40	0.04	1.180	33.8983	498.6441

After this transformation, item parameter estimates for grade 5 operational items were expressed in the scale score metric. Next, item parameters in the scale score metric for the grade 5 operational multiple-choice items were used as anchors to equate grade 4 and grade 6 linking itearap[pems administered to grade 5 students onto the base grade scale. In the next step, an anchor set was created for linking the grade 4 operational test to grade 5. This anchor set contained grade 4 items administered to grade 5 students and grade 5 items administered to grade 4 students. In the following step, the grade 4 operational test was calibrated and linked to the base grade, and a set of transformation constants was obtained that allowed placing grade 4 onto grade 5 scale. After the grade 4 operational test was placed on grade 5 scale, the grade 4 operational multiple-choice items were used as anchors to link the grade 3 items administered to grade 4 students to the grade 4 scale (same as grade 5 scale). The linking of the grade 3 operational test to the grade 4 scale was conducted via anchor set consisting of grade 4 items administered to grade 3 students and grade 3 items administered to grade 4 students. Linking of grades 6, 7 and 8 to the base grade was conducted in the same manner. After linking, all grades' obtained item parameter estimates on the same, content specific scale, and were expressed in the scale score metric.

Scale evaluation

The new vertical scales were evaluated in terms of their test characteristic curve (TCC) ordinality, scale statistical properties, and scale score distribution.

Test Characteristic Curves

The estimated item parameters on vertical scale were used to produce grade specific test characteristic curves (TCC) and standard error (SE) curves which allowed for visual evaluation of TCC ordinality reflecting increasing test difficulty across grades. Figures 1 and 2 show TCC and SE curves for vertically scaled grade 3-8 assessments in ELA and Mathematics, respectively.

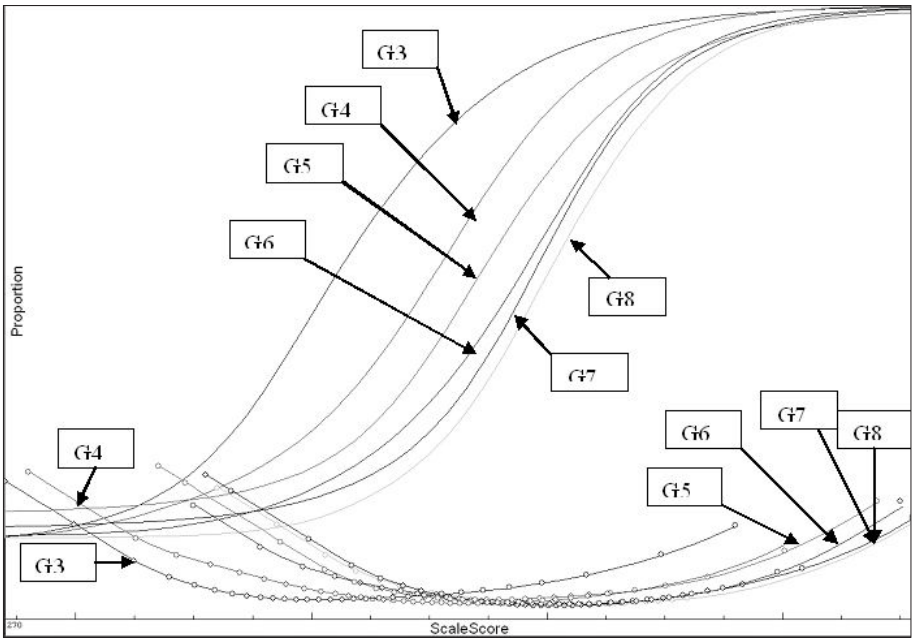

Figure 1. ELA grades 3-8 TCC and SE curves

Note that TCCs and SE curve for different grades are in different colors: blue-grade 3, pink-grade 4, green-grade 5, brown-grade 6, dark blue-grade 7, and lime-grade 8.

As seen in Figure 1, satisfactory ordinality was achieved for the ELA scale. The greatest TCC separation is observed between grades 3 through 6, and less separation is visible for grades 6 through 8. The SE curves for all grades are U-shaped with the smallest error in the middle of ability scale for each grade. The SE is expected to increase at the lower and upper end of each scale.

As shown in Figure 2 (below), good ordinality was achieved for the Mathematics scale as well. The greatest TCC separation is observed between grades 3 and 4, and grades 7 and 8. Grades 4, 5, 6 and 7 TCCs are closer to one another. The SE curves for all grades are in expected U-shape with less error in the middle of the scale and increased error at the lower and upper end of each scale.

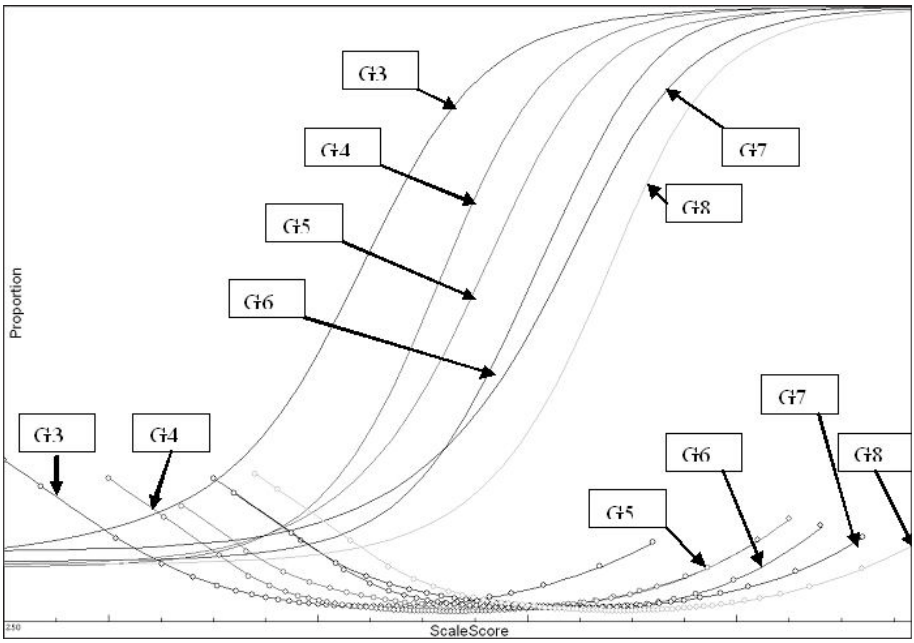

Figure 2. Mathematics grades 3-8 TCC and SE curves

Note that TCCs and SE curve for different grades are in different colors: blue-grade 3, pink-grade 4, green-grade 5, brown-grade 6, dark blue-grade 7, and lime-grade 8.

Taken together, the TCC's across grades for both ELA and Mathematics are parallel and ordinal indicating increasing test difficulty on the continuum. The grade 3 TCC is situated most to the left reflecting the fact that that grade 3 test is the easiest test and the grade 8 TCC is situated most to the right indicating the most difficult test of all 6 assessments in each content area. Uneven separation of TCC is often expected and reflects grade specific curriculum requirements and expectations.

Scale statistical properties

The number correct scoring method was employed to produce raw score to scale score conversion tables for each grade using inverse TCC method. The lowest and highest obtainable scores (called LOSS and HOSS) for each grade were set during the scoring table creation to reflect increasing values across grades. These scoring tables were used to score students and the group statistics were computed and are presented in Tables 2 (ELA) and Table 3 (Mathematics.)

Table 2. ELA grades 3-8 vertical scale properties

Grade level	Scale statistics		Mean difference between grades (in scale score points)	Scale range	
	Mean	SD		LOSS	HOSS
3	459.9	45.8		270	580
4	481.2	40.0	21.3	280	600
5	499.5	40.8	18.3	335	640
6	511.8	33.9	12.3	350	650
7	517.9	32.5	6.1	355	655
8	527.1	31.3	9.2	360	660

As presented in Table 2 and graphically illustrated in Figure 3, the ELA scale score means increase as grade level increases. The standard deviations illustrate a decreasing pattern and some ELA scale shrinkage. The lowest and highest obtainable scores increase as grade level increases. It should be noted that the mean difference between grades is not uniform across grade levels. This example is consistent with expectations that more growth is observed between lower grade levels (especially from grades 3 to 4 and from grades 4 to 5) and less growth is observed for higher grades (especially between grades 6 and 7 and between grades 7 and 8).

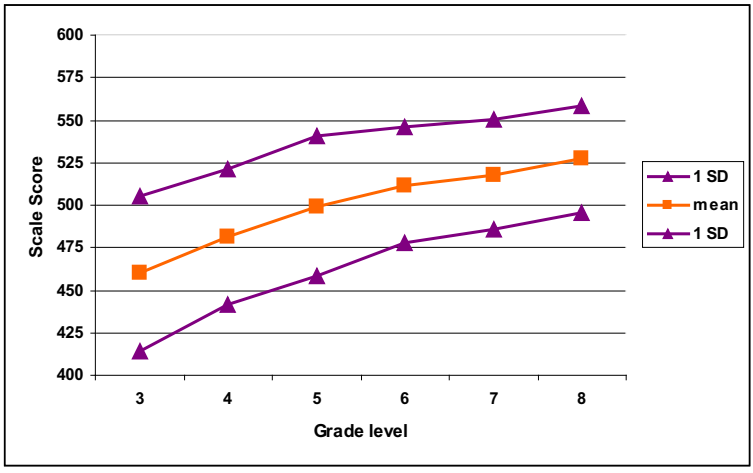

Figure 3. ELA grades 3-8 vertical scale properties

Table 3. Mathematics grades 3-8 vertical scale properties

Grade level	Scale statistics		Mean difference between grades (in scale score points)	Scale range	
	Mean	SD		LOSS	HOSS
3	463.4	41.1		250	560
4	487.3	38.0	23.9	300	610
5	500.0	39.7	12.7	335	625
6	511.6	41.9	11.6	350	640
7	522.5	39.7	10.9	360	660
8	541.1	46.7	18.6	370	690

As presented in Table 3 and illustrated in Figure 4, Mathematics scale score means also increase as grade level increases. The standard deviations do not show any specific trend and range from 38 scale score points for grade 4, to 46.7 points for grade 8. The grade 8 standard deviation is larger than the standard deviations for other grades. The lowest and highest obtainable scores increase as grade level increases. Similarly to observed ELA vertical scale properties, the mean difference between grades is not uniform across grade levels. Most growth is observed between grades 3 and 4 and less growth is observed between grades 4 through 7. Again, more growth is seen between grades 7 and 8.

It should be noted that in both scales (ELA and Mathematics) the grade-to-grade growth as reflected by scale score means was consistent with the TCC ordinality and separation.

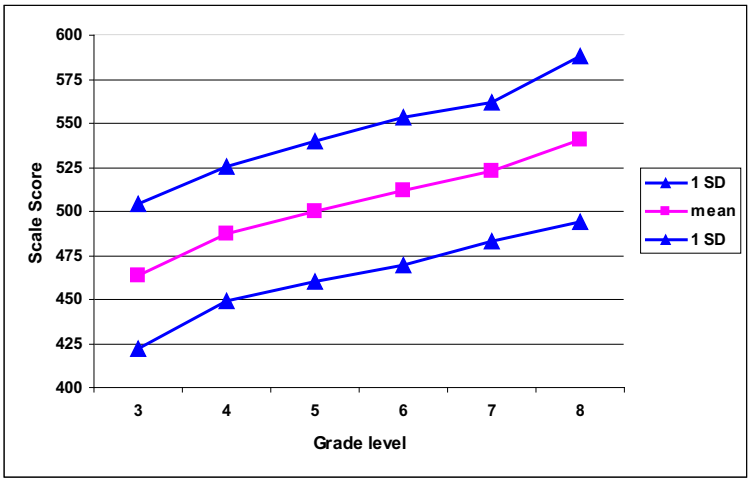

Figure 4. Mathematics grades 3-8 vertical scale properties

Scale score distribution

In addition to the evaluation of grade to grade growth using scale score mean changes across grades, the pattern of scale scores at the 10th, 25th, 50th, 75th, and 90th percentiles was examined across grades. Tables 4 and 5, below, summarize this information for ELA and Mathematics, respectively.

Table 4. ELA scale scores at different percentiles across grades

Grade	Percentile				
	10th	25th	50th	75th	90th
3	405	432	458	487	518
4	432	459	483	506	528
5	453	476	499	524	548
6	470	491	513	532	551
7	479	499	518	538	557
8	489	508	527	546	565

Table 5. Mathematics scale scores at different percentiles across grades

Grade	Percentile				
	10th	25th	50th	75th	90th
3	415	438	463	487	519
4	441	465	488	509	532
5	451	476	499	524	547
6	461	487	513	538	560
7	474	501	525	548	568
8	486	514	541	569	597

As expected and shown, in Tables 4 and 5, the scale score at the same percentile increase across grades for both ELA and Math, indicating a grade-to-grade growth at selected percentiles.

In addition, graphical representations of scale score cumulative frequencies were produced for both of the ELA and Mathematics scales. These are presented in Figures 5 and 6, respectively.

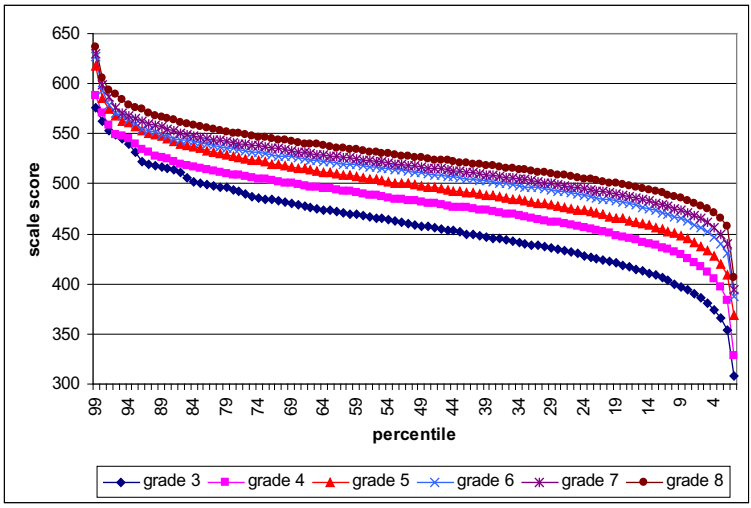

Figure 5. ELA scale score cumulative frequency distribution across grades

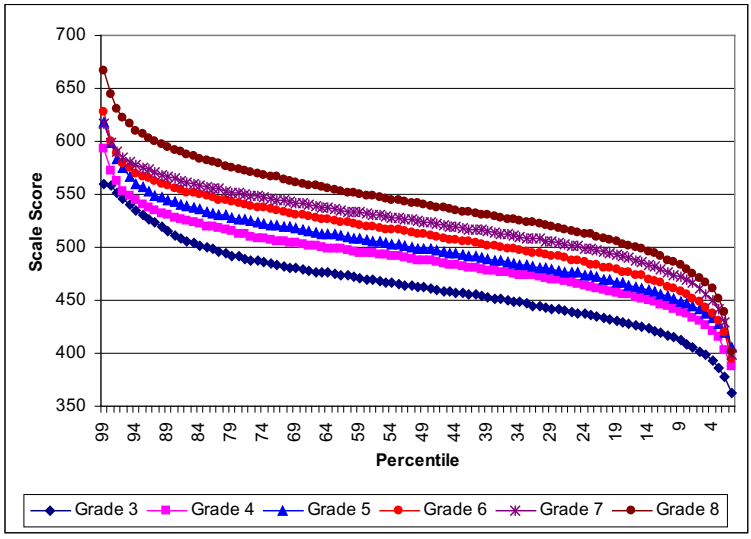

Figure 6. ***Math scale score cumulative frequency distribution across grades***

As indicated by Figures 5 and 6, both scales resulted in a reasonable separation of the grade distributions.

Summary of results

The vertical scales for both ELA and Mathematics yielded reasonable results. For both content areas sufficient and reasonable grade-to-grade growth patterns, in terms of mean scale score difference between adjacent grades, was achieved. A trend of standard deviations across grades was somewhat different for ELA and Math, but there is no expectation that the standard deviation pattern will be consistent across content areas. The scale score ranges were consistent across grades within each scale. In addition, scale score distributions at different percentile ranks were reasonable across both ELA and Mathematics scales. Taken together, it appears that the separate calibration method employed to develop vertical scale in ELA and Mathematics was appropriate given the data collection and test design. A variety of scale maintenance methods can be implemented in subsequent years to investigate and evaluate scale stability and grade-to-grade growth patterns resulting from these methods.

Practical application of vertical scales

Vertical scaling brings several valuable features to achievement tests. Vertical scales facilitate the estimation and tracking of growth over time (i.e., comparable scale scores) on individual students or groups using grade appropriate assessments. Two types of comparisons can be made: one grade to another in the same administration year or one cohort of students to another at any point in time (Patz, 2007.)

Vertical scaling of test forms also allows for important comparisons regarding test items and can lead to more efficient field testing of new content, as items targeted for one grade might be found to be of more appropriate difficulty for an adjacent grade, assuming that the targeted content standard is present in both grades. Final form selection for a target grade can then identify appropriate items from a larger pool, when each item in the pool has parameters on a common scale (Patz, 2007.)

In addition, grade specific proficiency standards can be set in a developmentally appropriate manner by considering not only knowledge, skills, and abilities measured by items in a grade specific test, but also knowledge, skills, and abilities measured by test items in adjacent grades. Setting progressive proficiency cut scores on a vertical scale will allow for meaningful and well articulated proficiency data trends on a learning continuum (Lewis & Haug, 2004.)

However, in addition to the benefits of a vertical scale, there are also several constraints underlying the interpretation of scale score changes from year to year. First, the standard error of measurement for an individual scale score tends to be rather large, especially for students at high and low ability levels. The intervals around the scale score for year (n) plus or minus one standard error and for year (n + 1) plus or minus one standard error will tend to overlap for many, perhaps most, students making it difficult to assess whether the difference in scale score between two years is a result of actual growth or can be attributed to the error of

measurement. Because the error of measurement tends to be random, aggregated scores will not have this problem for large enough case counts, and thus, group data may be compared from year to year with greater reliability.

Second, extreme scores tend to be affected by the statistical artifact referred to as, *regression to the mean*. This means that students who score very low in a given year will tend to have higher scores in the following year. Similarly, students with very high scores in a given year will tend to have somewhat lower scores the following year. This decreases the ability to interpret score growth from one year to the next using vertical scaling.

In addition, maintenance of a vertical scale can be quite complex. A common item method is considered to be most effective to maintain a vertical scale from year to year. In this method, common items across administrations are used to link new tests to the previous year tests. If no common items are administered from year to year, it may be necessary to administer off-grade level items (in addition to regular on-grade level items) every year to maintain the scale.

Both, the benefits and limitations of a vertical scale need to be carefully considered before making a decision of using such a scale in a testing program. First, in order for educators to benefit from a vertical scale, the scale needs to be carefully designed and constructed, and a number of conditions must be met. The most important one is having vertically aligned learning standards with considerable grade-to-grade overlap and a systematic, intentional increase in form grade-to-grade difficulty. Next, an appropriate vertical scaling design including the psychometric procedures, data collection plan, and appropriate numbers of common items across levels or numbers of students taking multiple forms, needs to be specified. It is very important that the data collected to create a vertical scale be gathered under conditions closely approximating the operational conditions and that data consists of large, statistically representative samples of students (Patz, 2007). Finally, a policy decision must be made in regard to use of the test scores. Are the scores primarily used to classify students into proficiency levels in a given grade level? Are they used to determine individual students' progress as they move from grade to grade? Are they used to determine eligibility of students for additional instruction time? Are scores used for promotion to the next grade? Are they used for tracking cohort performance over time? Are they used as indicators of teacher performance or in school evaluation? These and other questions must be addressed in a broad context of the testing program including test stakes and consequences of test results to students and teachers.

Despite these limitations and policy considerations, when vertical scales are well defined and constructed for use in large-scale educational testing programs, they can significantly enrich the interpretations of test scores by providing a systematic way to examine progress of student knowledge and skill acquisition in respect to implemented curricula across grade spans.

Future possibilities in Poland

Given that the current testing program in Poland includes assessments in grades 6, 9 and high school exit exam (equivalent of grade 12) it would be extremely difficult to develop a meaningful vertical scale spanning these grades. The National Assessment of Educational Progress (NAEP) in the United States attempted to vertically link forms for grades 4, 8, and 12, but abandoned the effort because comparisons of students separated by 4 or 8 years were too difficult to interpret (Haertel, 1991.) It would seem likely that in these applications the most fundamental vertical scale requirement - overlap in test content and learning standards for levels to be linked was not satisfied. The Polish educational testing program would likely face similar challenges. If there is ever a need or desire to link grade specific assessments across grades and construct a vertical scale in Poland, it is recommended that the assessments be administered in consecutive grades and careful consideration be given to test content similarity and overlap in adjacent grades.

References

1. Gao, F., & Chen, L. (2005). Bayesian or Non-Bayesian: A Comparison Study of Item Parameter Estimation in the Three-Parameter Logistic Model. Applied Measurement in Education, 18 (4), 351-380.
2. Haertel, E. (1991). *Report on TRP analyses of issues concerning within-age versus crossage scales for the National Assessment of Educational Progress* [ERIC Clearinghouse Document Reproduction Service No ED404367]. Washington, DC: National Center for Education Statistics.
3. Huynh, H. & Schneider, C. (2004). Vertically Moderated Standards as an Alternative to Vertical Scaling: Assumptions, Practices, and an Odyssey through NAEP. Paper presented at the National Conference on Large-Scale Assessment, June 21, 2004, Boston, MA.
4. Kolen, M. J. & Brennan R. L. (1995). Test equating. Methods and practices. New York, NY: Springer-Verlag.
5. Lewis, D. M., & Haug, C. (1994) A Standard Setting Odyssey: On a quest for across-grade consistency. Paper presented the annual meeting of the National Council on Measurement in Education, April 2004, San Diego, CA.
6. Lissitz, R. & Huynh, H. (2003). Vertical equating for the Arkansas ACTAAP assessments: Issues and solutions in determination of adequate yearly progress. A report submitted to the Arkansas Department of Education.
7. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*.
8. Hillsdale, NJ: Erlbaum.
9. Mislevy, R. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.
10. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
11. No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
12. Patz, R.J. (2007) Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems. Paper Prepared for the Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.
13. Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.
14. Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.