computers in the teaching process

Computer tools in education, 2018
№ 5: 24–40
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2018-5-24-40

PROCESSING AND VISUALIZATION OF TEST-RESULTS DATA

Monakhov V. V. [1], Kozhedub A. V.[1], Khannanov N. K.[2], Korolev A. A.[3], Kurashova S. A.[3]

[1]Saint Petersburg State University, Saint Petersburg, Russia
[2]Autonomous nonprofit general education organization "New Chernogolovskaya School", Chernogolovka Russia
[3]Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Saint Petersburg, Russia

Abstract

A data processing method is proposed for estimating the measurement error of scaled test scores; this involves scaling half-test raw scores to effective full-test raw scores and converting them to scaled test scores. It is shown that the method allows to estimate the measurement error with high accuracy.

The proposed approach to data visualization involves the removal of the part of the data corresponding to the rarest values. In the case of very noisy data, this method helps to eliminate the contribution of atypical values and provides a significant increase in the visibility of scatter plots.

Keywords: *data processing, data visualization, computer aided assessment.*

Citation: V. V. Monakhov, A. V. Kozhedub, N. K. Khannanov, A. A. Korolev, and S. A. Kurashova, "Processing and Visualization of Test-Results Data," *Computer tools in education*, no. 5, pp. 24–40, 2018. doi:10.32603/2071-2340-2018-5-24-40

In order to estimate conditional standard error of measurement (CSEM) it is common to re-test the same students using the same test and regard two testing occasions as parallel test forms (half-tests). There is a chance, however, that students could have memorized some of the tasks, gained extra knowledge or vice versa forgotten certain things during the time that has passed between the two testing sessions. In addition, in many cases retesting is impossible or undesirable. That is why it is preferable to split the test into two item sets (half-tests) that would be regarded as parallel test forms.

However, such parallelism might not be ideal, and the difference between results of half-tests may be significant. Already developed test cannot be split into pairs of items of the same difficulty because it is virtually impossible to predict precise difficulty of items for an unknown group of testees, especially when used in computer aided assessment. That is why it is necessary first to develop a methodology for splitting the items into two almost parallel test forms and compensating for non-parallelism. However, it is not enough to have two parallel half-tests that allow estimating error of measurement of the raw scores because the standardization procedure (converting raw scores to scaled scores) is not linear and may be applied only to full-test raw scores. This means that the estimated error of measurement of the raw scores cannot be transformed directly into error of measurement of the scaled scores.

This paper proposes a method to estimate error of measurement of scaled scores; this approach was used to estimate error of measurement of the Russian Unified State Exam (USE) in physics taken by 10,472 students in Moscow in 2010. Data processing software was developed on the basis of BARSIC software [5]. In this paper the set of items being studied will be called a test, although it may be a set of problems or other tasks that do not take a form of a test.

1. METHODS OF ESTIMATING SCORE LEVEL STANDARD ERRORS

Based on the results of parallel test forms it is possible to estimate the standard error of measurement S_E [1, 8, 11]:

$$S_E = S_X(1 - r_{xx'})^{1/2}, \tag{1}$$

where S_X is the standard deviation (mean squared variation) of the test scores, $r_{xx'}$ is the correlation coefficient between the two parallel test forms. It is important to note, however, that formula (1) is only applicable when the scores are normally distributed; otherwise it does not make sense [1].

Existing publications [1, 8, 11] list several methods to estimate conditional error of measurement.

Following the classical test theory model, Thorndike's method [10] uses equation (2) to estimate S_E:

$$S_E = S_{X_1 - X_2}, \tag{2}$$

where $S_{X_1-X_2}$ is mean squared variation of the difference between X_1 and X_2 scores in parallel half-tests; mean values of X_1 and X_2 for each point (let's denote them $< X_1 >$ and $< X_2 >$) are regarded exactly equal. Normally, specific value intervals (score levels) are considered. The main drawback of the method is low accuracy for small numbers of testees. However, the method is very stable and its accuracy is known from the classical theory of measurement.

Polynomial method is a refinement of the Thorndike's approach. The dependence of $(X_1 - X_2)^2$ from X_1 scores is approximated by $Y(X_1)$ polynomial and is regarded as the value of the error at the point X_1. Normally, the prediction would be made using cubic or fourth degree polynomial. However, this approach should only be used when the distribution shape is known.

Lord's binomial error model [2] regarded the test as a randomly selected set of k items. The prediction was not very accurate and Keats [3] improved the formula. Lord's compound binomial approach was a refinement of the binomial error model and is based on the selection of stratified samples of items (subtests of different item categories) rather than completely random samples — Lord's binomial error model is then applied to each stratified sample. Formulae derived as part of these methods can only be regarded as some basic approximations unrelated to a particular participant, group of participants or even item set. Theoretical accuracy or adequacy of these theoretical models remains unknown.

Other estimates [8, 11] based on Rasch model [9] consider parameters of individual items and estimate deviations from theoretical values that were to be obtained in a test with given characteristics. Before using these methods it should be proven that for all items the necessary conditions for Rasch theory are fulfilled, which is very often not the case.

In this paper to estimate conditional error of measurement of raw scores we used Thorndike's method and deviations from a straight line approximating the relationship between raw scores for one half-test and scores for the other half-test. Large sample size allowed obtaining highly accurate estimates. However, this method cannot be used to estimate error

of measurement of scaled test scores. We have proposed a method to convert raw half-test scores to scaled test scores and estimate error of measurement with high accuracy.

2. ESTIMATING ERROR OF MEASUREMENT OF RAW RUSSIAN USE SCORES

A relatively high number of items in Russian USE in physics allowed forming two item sets (half-tests) and regarding them as parallel tests.

Fig. 1 shows relationship between raw scores for even items and raw scores for odd items (scatter plot). For greater clarity about 5 % of the results providing the greatest scattering are not included in the picture. The method used to perform this will be discussed further. The actual data analysis included all values.

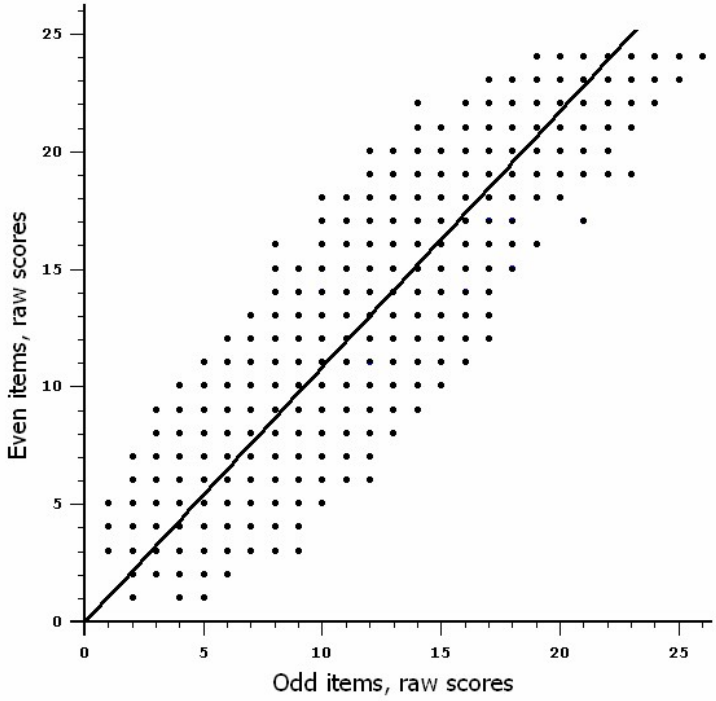


Figure 1. Relationship between raw scores for even items and raw scores for odd items

Parameters of linear approximation

$$y = ax + b, \tag{3}$$

were calculated using ordinary least products regression (OLP) [4]:

$$a = 1.07,\ b = 0.05,\ w_x = 2.87,\ w_y = 3.06,\ R = 0.851, \tag{4}$$

where w_x is mean square deviation of X from the regression line, w_y is mean square deviation of Y from the regression line, and R is correlation coefficient. The scatter plot (fig. 1) shows that the visible scattering is somewhat higher for the average scores. Accordingly, the standard error of measurement is greater in this range and is decreasing towards higher and lower score ranges, which is consistent with the results of similar research [1, 8].

The item sets used can be considered parallel but having a slight difference in the regression coefficient ($a = 1.07$).

Deviations from the regression line conformed to normal distribution. It is natural to assume that the error of measurement for even and odd items are independent and that the standard error of measurement of the scores on the scaled $y' = y/a$ axis (even items) coincide with the error of measurement of the x axis (odd items) and will be denoted by ε.

Let's denote the true scores (exact values that participants would have obtained in the absence of error of measurement) by x and y and observed scores by $\tilde{x}$ and $\tilde{y}$.

In this case:

$$\tilde{x} = x + \tilde{\varepsilon}_x, \tag{5}$$

$$\tilde{y} = y + \tilde{\varepsilon}_y, \tag{6}$$

where $\tilde{\varepsilon}_x$ and $\tilde{\varepsilon}_y$ are errors.

We'll use the triangular brackets to denote mean squared values. Then:

$$< \tilde{\varepsilon}_y >= a\varepsilon,$$

$$< \tilde{\varepsilon}_x >= \varepsilon.$$

Because of the large number of measurements the error related to the approximation was small and did not exceed 0.06 score at 95 % confidence level. Therefore we can assume that the regression line (3) gives the exact values, i.e. that $y = ax + b$, and from (3) and (6):

$$\tilde{y} = ax + b + \tilde{\varepsilon}_y. \tag{7}$$

Using (5) and (7) we conclude that:

$$\tilde{y} = a\tilde{x} + b + (\tilde{\varepsilon}_y - a\tilde{\varepsilon}_x).$$

This means that $\Delta\tilde{y}$ (deviation of the observed score $\tilde{y}$) from the $a\tilde{x} + b$ regression line equals:

$$\Delta\tilde{y} = \tilde{\varepsilon}_y - a\tilde{\varepsilon}_x.$$

This is different from error $\tilde{\varepsilon}_y$ because x axis depicts observed scores $\tilde{x}$ rather than true scores x. Therefore $(\tilde{x}, \tilde{y})$ has error of measurement both on the x and y axes.

In accordance with the assumptions made:

$$< \Delta\tilde{y} >=< \tilde{\varepsilon}_y - a\tilde{\varepsilon}_x >= \sqrt{< \tilde{\varepsilon}_y >^2 + a^2 < \tilde{\varepsilon}_x >^2} = \sqrt{a^2\varepsilon^2 + a^2\varepsilon^2} = \sqrt{2}a\varepsilon.$$

In other words, the mean square deviation of y from the regression line ($w_y =< \Delta\tilde{y} >$) is $\sqrt{2}$ times greater than standard error of y and $a\sqrt{2}$ times greater than standard error $< \tilde{\varepsilon}_x >$ of x.

Total observed full-test raw score equals:

$$\tilde{x} + \tilde{y} = x + y + (\tilde{\varepsilon}_x + \tilde{\varepsilon}_y).$$

This means that error of the total raw score $\tilde{\varepsilon}$ equals:

$$\tilde{\varepsilon} = \tilde{\varepsilon}_x + \tilde{\varepsilon}_y.$$

Therefore:

$$< \tilde{\varepsilon} >=< \tilde{\varepsilon}_x + \tilde{\varepsilon}_y >= \sqrt{\varepsilon^2 + a^2\varepsilon^2} = \varepsilon\sqrt{1 + a^2}.$$

Let's denote $< \tilde{\varepsilon} >$ by σ. As a result we get:

$$\sigma = w_y \sqrt{\frac{1+a^2}{2a^2}}. \tag{8}$$

In accordance with (4) and (8) standard error of Russian USE raw scores can be estimated as $\sigma = 3.0$. As we can see (Fig. 1), the scattering at the ends of the range is somewhat smaller than in the middle of the range. This fact is consistent with the tendencies described in other works [1, 8, 12].

3. CONVERTING RAW SCORES TO SCALED SCORES

Unfortunately, raw scores for a half-test cannot be easily scaled. Only full-test raw scores may be converted to full-test scaled scores by means of non-linear scaling function. So we propose a two-stage scaling method:

I) linear scaling of half-test raw scores to pseudo-full-test effective raw scores;

II) scaling of effective raw scores to scaled effective scores by means of standard non-linear scaling function.

We can assume that if the half-test raw scores for odd or even items are multiplied by a scaling factor that ensures the maximum of 50 raw scores (maximum for full-test raw scores scale) then after scaling of such pseudo-full-test raw scores the resulting distribution of scaled pseudo-full-test scores can be analyzed in the same manner as we have done for the raw scores. However, this method is correct only if a is equal to 1. Also when scaling half-test raw scores to pseudo-full-test raw scores the error would be $\sqrt{2}$ times greater than for the true full-test raw scores because when adding the results for odd and even items standard error will not double but increase by $\sqrt{2}$ times. That is why the resulting scaled score distribution will be incorrect. To obtain the correct distribution in the first stage of scaling it is necessary to scale the scores so that a equals 1 and decrease the distance between the points and the regression line in $\sqrt{2}$ times. We scale scores on the axes so that after the conversion a coefficient for the regression line equals 1. In our case maximal score for the first half-test (24 scores for even items) is to become maximal raw score (50 scores). Maximal score for the second half-test (26 scores for odd items) will be somewhat greater than maximal raw score. Therefore, after the conversion raw scores exceeding the allowable maximum must be replaced by the maximum allowable — otherwise it is impossible to convert them to scaled test scores. This may seem doubtful, but firstly the number of participants with the highest scores is so low that any operations with the scores "going off-scale" have almost no effect on the overall results. Secondly, such "off-scaling" takes place in the reality (very high-scoring subjects cannot get any higher than the maximum, which is 50 raw scores), so the replacement of the off-scale scores by the maximal scores is quite correct.

Scores obtained through this procedure will be called “effective full-test raw scores” where “effective” denote that these are approximated scores that provide scaling of half-test results to entire test score rather than real raw scores.

It should be noted that after scaling the scores will no longer be integer, but our analysis showed that they should not be rounded as it introduces an additional error.

If we assume that in (3) $b = 0$ (taking into account the dispersion of several scores b of 0.03 score can be neglected) we obtain the following equation to convert the point with coordinates $\tilde{x}$ and $\tilde{y}$, which reduces the distance to the regression line $y = ax$ in $\sqrt{2}$ times and gives the point with coordinates $\tilde{x}$ and $\tilde{y}$:

$$\tilde{x}' = \frac{(1+\frac{1}{\sqrt{2}})\tilde{x} + (1-\frac{1}{\sqrt{2}})\frac{\tilde{y}}{a}}{2}. \tag{9}$$

$$\tilde{y}' = \frac{(1-\frac{1}{\sqrt{2}})\tilde{x} + (1+\frac{1}{\sqrt{2}})\frac{\tilde{y}}{a}}{2}. \tag{10}$$

The approach outlined allows us not only to evaluate the average error for the entire dataset but also makes it possible to measure the error of measurement for single intervals and track deviations from the regression line for individual participants as well as systematic deviations for groups of participants, that is, measure the conditional error of measurement.

4. SCATTER PLOTS AND ERROR OF MEASUREMENT OF THE RUSSIAN USE SCORES

Figure 2 shows the scatter plot obtained by processing the data from figure1 through (9)–(10) formulas.

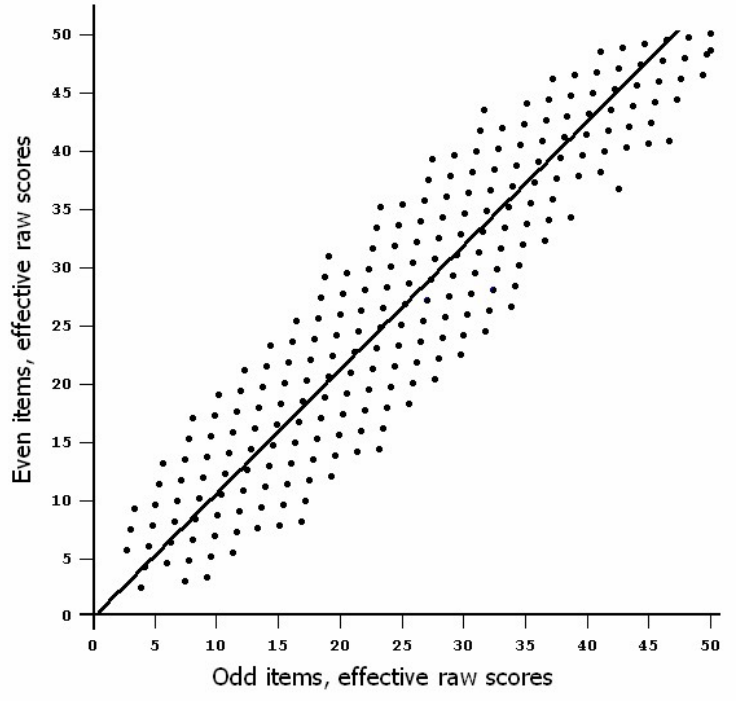


Figure 2. Relationship between effective full-test raw scores for even items and effective full-test raw scores for odd items

Figure 3 shows similar dependence for the same data converted to effective full-test scaled scores on a 0 to 100 scale — we will call them “effective” scores to stress that these are scaled scores from a half-test.

The value of the standard deviation w for the entire range was 5.7 scaled scores. Under the assumption of equality of standard error for even and odd items, their independence and normal distribution, the standard error for each item set is $\sqrt{2}$ times smaller, i.e. $\sigma = 4.0$ scaled scores.

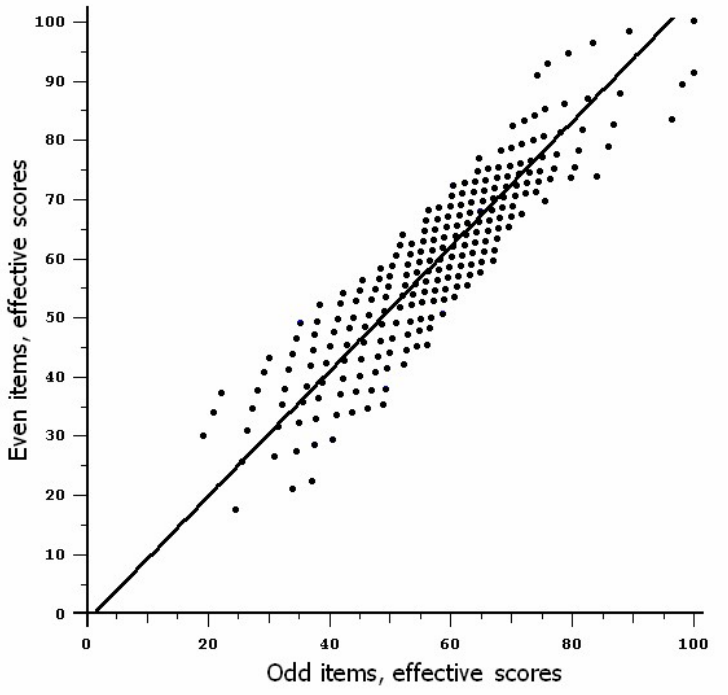


Figure 3. Relationship between effective scaled scores for even items and effective scaled scores for odd items

But for points in a linear scaling range (from 53 to 75 scaled scores) the deviation for odd items was less and was $w = 4.74$ scaled scores, i.e. $\sigma = 3.35$ scaled scores.

Outside the linear scaling range of conversion (less than 53 or more than 75 scaled scores) the error increases which can be clearly seen in Figure 3 as an increased dispersion of the points. For the range of 21 to 53 scores $w = 6.0$ scaled scores and the standard error is $\sqrt{2}$ times less and equals to $\sigma = 4.24$ scores. This is of course the average error within the specified range. For the range of 75 to 100 scores $w = 7.95$ scaled scores and the standard error is $\sigma = 5.6$ scaled scores.

The distortions associated with the scaling procedure can be estimated by comparing the half-sum of effective scaled scores for the half-tests with the real scaled score for the entire test. Ideally, if the errors were normally distributed and there were no distortions (due to the boundaries of ranges or effects of second order) these values should have been equal. It was found that standard error introduced due to the scaling procedure was 0.28 scores for the 21–52 scaled score range, 0.11 for the 53–75 scaled score range and 0.55 for the 76–100 scaled score range. These values are small compared to error values for these ranges (less than 10 %), which indicates correctness of the proposed method of scaling.

5. COMPARISON OF RUSSIAN USE RESULTS IN MATHEMATICS AND PHYSICS

In another paper [7] we tried to estimate error of measurement of the Russian USE scores in Physics. It was 7.0 scores based on Rasch model and 6.1 ± 0.6 scores based on the comparison of scores in math and in physics. The closeness of these values allows assuming that the dispersion on the scatter plot of the relationship between scores in physics and scores in mathematics is almost completely determined by measurement errors. We have shown [7] that presence of statistical variation prevents accurate differentiation of abilities of high scoring participants

(attaining 75 to 100 scaled scores for 100-score scale) who are the most interesting group for the leading universities. Those abilities can be more accurately measured by Science Olympiads (such as Online Olympiad in physics described in [6] because Olympiad tasks are much more difficult than USE items.

The results of the current study indicate that our estimate of Russian USE score in physics (Monakhov, 2011) was correct for the high score range (75–100 scores). That is why the conclusions about the limited applicability of the Russian USE for the measurement of abilities of University applicants stand good.

At the same time, direct estimates of the error using half-test approach showed that on average it is equal to 4.0 test scores and for average score range it is 3.35 scores, which is about two times smaller than the error found in [7] using Rasch model. This can be explained by the fact that in [7] the estimate did not take scaling procedure into account and therefore the results should be treated as related to raw scores rather than scaled scores. The 0 to 100 scaled score range used in the model should be replaced with a 0 to 50 raw scores range. Because the score range is reduced by 2 times the error estimate obtained in [7] should also be reduced by a factor of 2. That is, the standard error calculated using Rasch model is equal to 3.5 raw scores, which is satisfactory consistent with the value of 3.0 raw scores obtained in the current investigation.

Estimated error of measurement of physics and math test scaled scores of 6.1 that was obtained in [7] is an upper bound. It included both scatter of points due to the statistical nature of the correct answers and the scatter arising from differences in subject areas in the exams in physics and mathematics. The results of this study enable us to refine the conclusions of [7]. If we assume that the standard error of measurement in mathematics is the same as the error in physics and equal to 4.0 scaled scores then the mean squared contribution from dispersion due to the difference in subject areas can be estimated by 2.3 scores. Thus, although it is non-zero, but its scatter effect is generally insignificant (without it the variance on the scatter plot of the results in physics and mathematics would be 5.7 scores instead of 6.1 scores).

6. DATA VISUALIZATION FOR ITEM SETS OF VARYING DIFFICULTY

It should be noted that the choice of odd and even items as a criterion for compilation of parallel test versions was determined by the fact that the resulting test versions had very similar difficulty. Therefore, the regression line in Figure 1 is almost passing through the origin and the slope of regression line is close to 1. However, this can only be the case for the item sets that have similar difficulty across all item difficulty range.

Figure 4 depicts the relationship between raw Russian USE total scores for item sets (parts) B and C and scores for part A. It seems to be linear with broad scattering band. However, this point of view is misleading.

B and C items are much more difficult than A items although total raw score for part A is 25 and maximal total score for B and C parts is also 25. It is obvious that the regression line is very distant from the origin. At the same time variance ($w_x = 3.5$ and $w_y = 4.2$) is much greater than in the first case (for the relationship between scores for even items and scores for odd items $w_x = 2.7$, $w_y = 2.9$ — see Fig. 1).

Moreover, if we remove the points overlapping less than 10 times (i.e., remove 535 points from the original dataset, which is about 5 % of all points), we obtain the picture shown in Fig. 5.

It is evident that the relationship is not linear and the width of the dispersion band strongly depends on the point at which the measurement is made. This fact was confirmed in the study of item sets of varying difficulty. Thus, it is fundamentally wrong to approximate the dependence

of the raw scores for item sets of varying difficulty with a linear dependence and the average measurement error in this case does not really characterize the actual measurement error. This is consistent with the results obtained in other works [1].

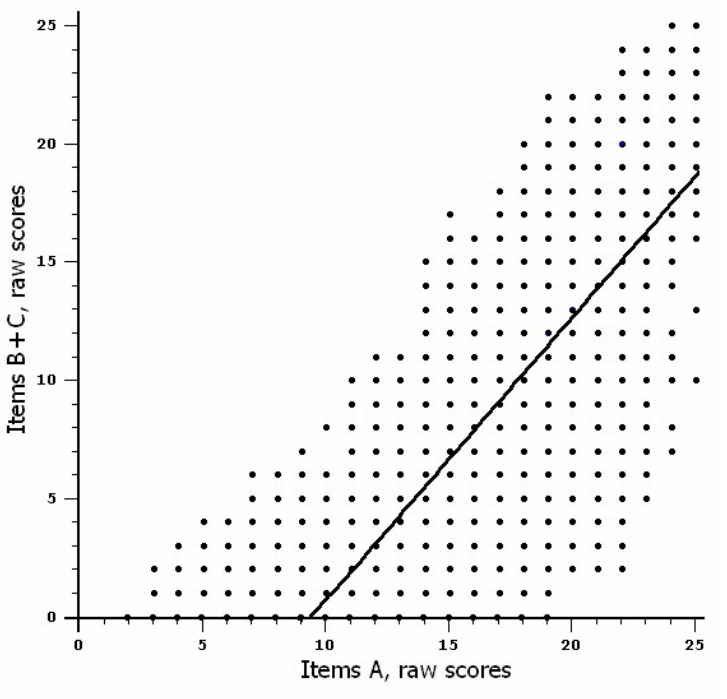


Figure 4. Relationship between total scores for parts B & C and total scores for part A

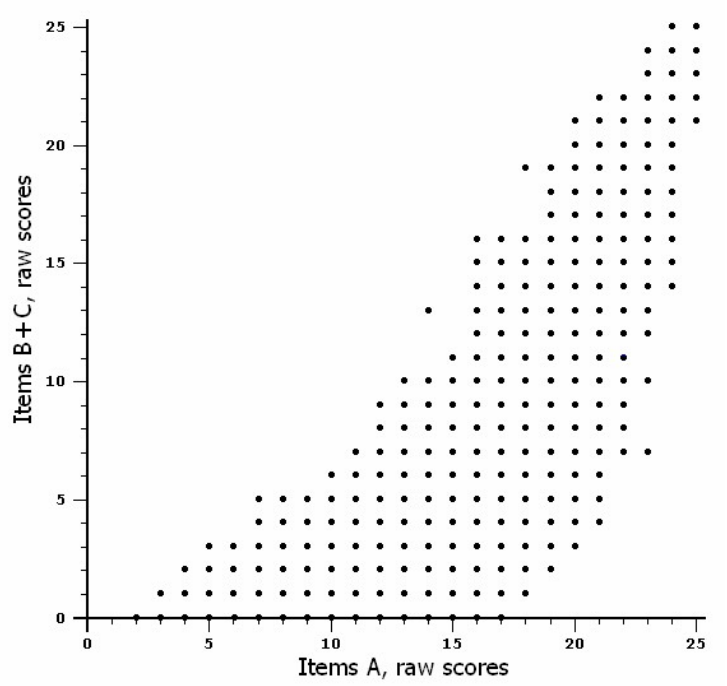


Figure 5. Relationship between total scores for parts B & C and total scores for part A after removal of 5 % of points (overlapping less than 10 times)

7. REMOVAL OF ATYPICAL DATA POINTS

Let us have a closer look at the proposed method of removing of some data records corresponding to the most atypical values. Usually, the data points distant from the core set of points or from a given line segment can be removed from the dataset. However, when there is a very noisy set of measurements, as in our case, this approach is ineffective, i.e. it is unclear how to use it to remove unnecessary data shown in Fig. 4. Our method is based on the analysis of frequencies of points in the nodes of a discrete two-dimensional grid. Each hit to the node adds one score to the point counter for the node. After going through all values in the original dataset those points that fall into the nodes with a count below the threshold should be removed. For example, if we set the threshold value of the counter to 5, only points of the grid nodes which got at least five points will remain in the filtered data set.

In this case the scores for two item sets are integers. However, there are situations where the coordinates of points (the values plotted on the axes) are real numbers or when the grid spacing (equals to 1) is too small. In this case, you can set the grid spacing along each axis in accordance with the requirements of a specific task and get points in the node set off on the coordinate value rounded to the nearest grid point.

Let us explain how this approach helps to improve the view of the scatter plots and other diagrams in case of very noisy data sets. Let us assume that σ is a standard deviation. Because of the statistical variance the probability of finding a point with coordinate $x, 3\sigma$ or more away from the mean position x_0, in the case of normal distribution is equal to 0.28 %. Usually such probability is ignored or else such points are discarded. If there are 10,000 points, then we get about 28 points that lie beyond the 3σ and distort the picture.

If the magnitude of change is comparable to the value of σ, as in our case in Figure 1 or especially in Figure 4, a value of 3σ is too large and we would like to see only the points that are not further away from the mean than 2σ (about 95 % of points) or even not further away than σ (about 68 % of points). But how can we remove extra points when we do not know the form of the dependence (Fig. 4)?

If there are 10,000 points, then we have about 500 points on the plot that lie beyond 2σ and about 3200 points that go beyond σ. It is obvious that near the mean value a point with a certain coordinate will get hundreds of dots and outside 3σ there will be single points but on the scatter plot they will all look the same. Therefore, the figure will depict the score distribution inadequately.

We suggest removing the points that have few hits in the same place of the scatter plots — we will call them atypical. Removing atypical points allows seeing the dependence and change in variation in cases where the original plot (without removing atypical points) was to a great extent filled with atypical points.

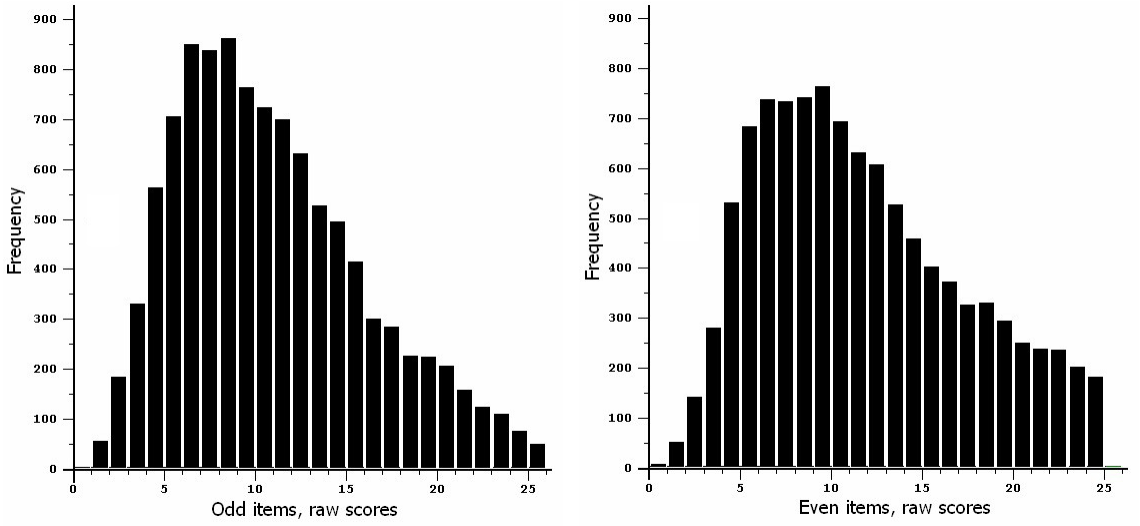
This method goes beyond the analysis of the Russian USE results and may be used for visualization of a variety of very noisy data sets. In addition, it can be used for removal (filtering) of atypical data in data processing.

8. TEST SCORE FREQUENCIES

Let us have a look at the histograms of test scores (Fig. 6) for odd and even item sets.

These graphs correspond to the initial data set shown in Fig. 1 (before filtering). You can see that these distributions are different and at first glance it seems that the half-test forms are not parallel.

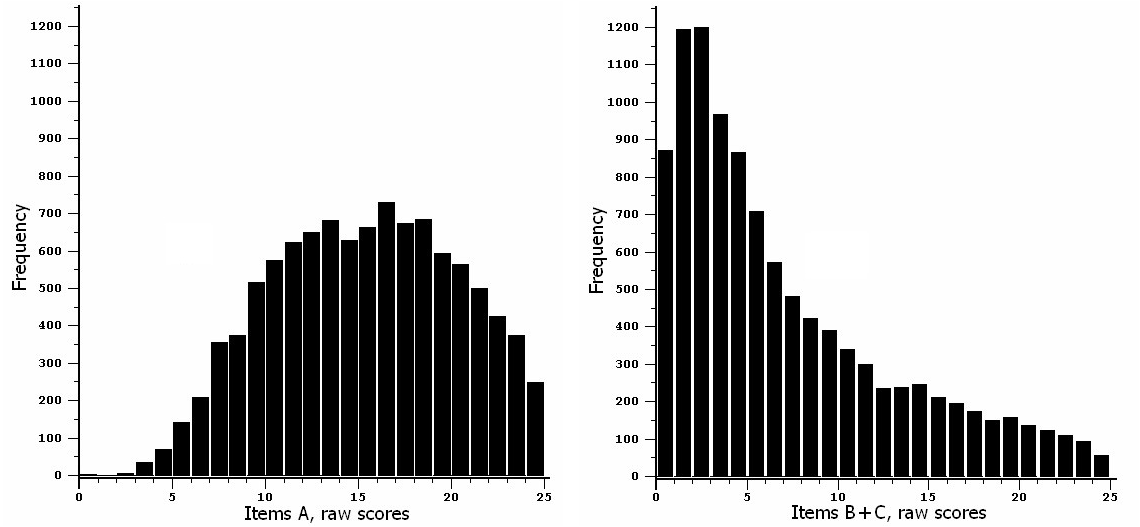
Of course, they are not ideally parallel but the situation is not as bad as you might think. The position of the peak on both bell-shaped curves in Fig. 6a and Fig. 6b is in the same area of 7.5–8.5 scores, and the percentage of correct answers differs by only 7 % being 40 % in the first case and 47 % in the second case.



(a) **(b)**

Figure 6. Test score frequencies for odd (a) and even items (b)

Figure 7 shows a histogram of the distribution of test score frequencies (i.e. number of participants who have obtained a specific score) for part A and parts B&C items. These distributions differ markedly from each other causing nonlinear relationship in Figure 5.



(a) **(b)**

Figure 7. Test score frequencies for part A (a) and parts B&C (b)

Half-test forms, one of which consists of part A items and another one of parts B and C items cannot be regarded as parallel. Difficulty of the first set is much lower than that of the second set, as indicated by the position of the peak in the bell-shaped function in Figure 7a (it is

shifted toward higher scores) and high percentage of correct answers (60 %). For the second set the peak in the dependence in Figure 7b is shifted toward lower scores and the percentage of correct answers is two times lower (27 %).

If both sets were parallel, the sum score of the two sets would have lead to a doubling of both scores and the height of the columns. The shape of the distribution would not change. Figure 8 represents the distribution of score frequencies for the full set of items (parts A+B+C).

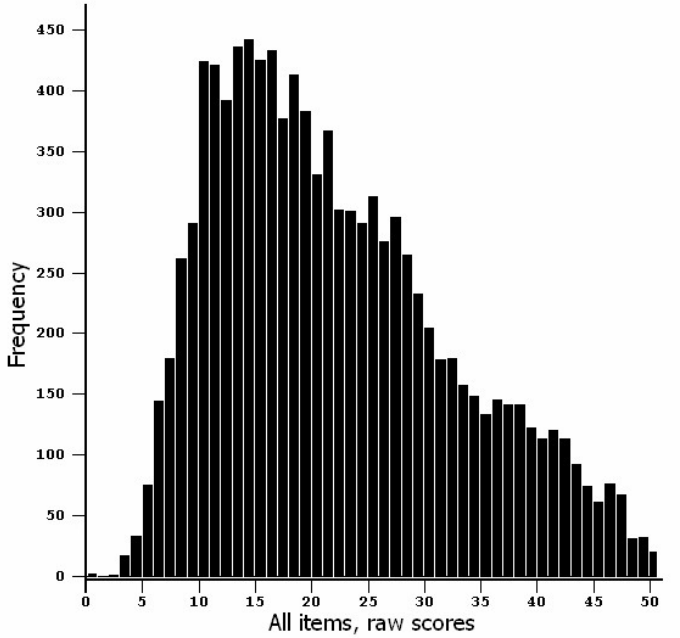


Figure 8. Test score frequencies for the entire item set

It is obvious that the distributions shown in Figure 7 have very different shape from the distribution for all items (Fig. 8) and that distributions in Figure 6 are much closer to the one in Figure 8. Odd and even items are usually specifically combined as parallel test forms when full test is being developed. So the fact that results of approximation (3), (4) are good is not so surprising. However, as we have found parallelism is not very good in this case.

9. SPLITTING ITEMS INTO GOOD PARALLEL TEST FORMS

Thus, for the correct split of the item set it is important to obtain similar distributions in the histograms and hence the symmetric distribution of points on the scatter plot depicting relationship between scores for the first half-test form and from scores for the second one (Fig. 1) as well as the same percentage of correct answers. The shape of the distributions is more important than the exact half-test scores obtained or the number of items in a set. At the same time the slope of the regression line (a should be close to 1) and its shift relative to the origin (b should be close to 0) might serve as good criteria. For example, when splitting items into even and odd ones maximum score for odd items was 26 and maximum score for even items was 24 but the sets could still be regarded as parallel.

To obtain the correct distribution of the results it is natural to include pairs of items of the same difficulty in the test. However, already developed test cannot be split into pairs of items of the same difficulty, as was mentioned above. And if you sorted the items by increasing difficulty

and selected sets from the pairs of adjacent items the first set would be always easier than the second one. To solve this problem we propose to alternate the order of the items selected for the sets from such pairs: take the first item from the first pair into set 1 and the second item from the first pair into set 2; take the second item from the second pair into set 1 and the first item from the second pair into set 2; take the first item from the third pair into set 1 and the second item into set 1; take the second item form the fourth pair into set 1 and so on. It is also possible to fine-tune the distributions by exchanging the items in sets 1 and 2 but it may be necessary only if corresponding items in different sets differ by the number of scores assigned to them or when parallel forms have different number of items.

Figure 9 shows the relationships between test scores for the parallel forms that were comprised using proposed approach.

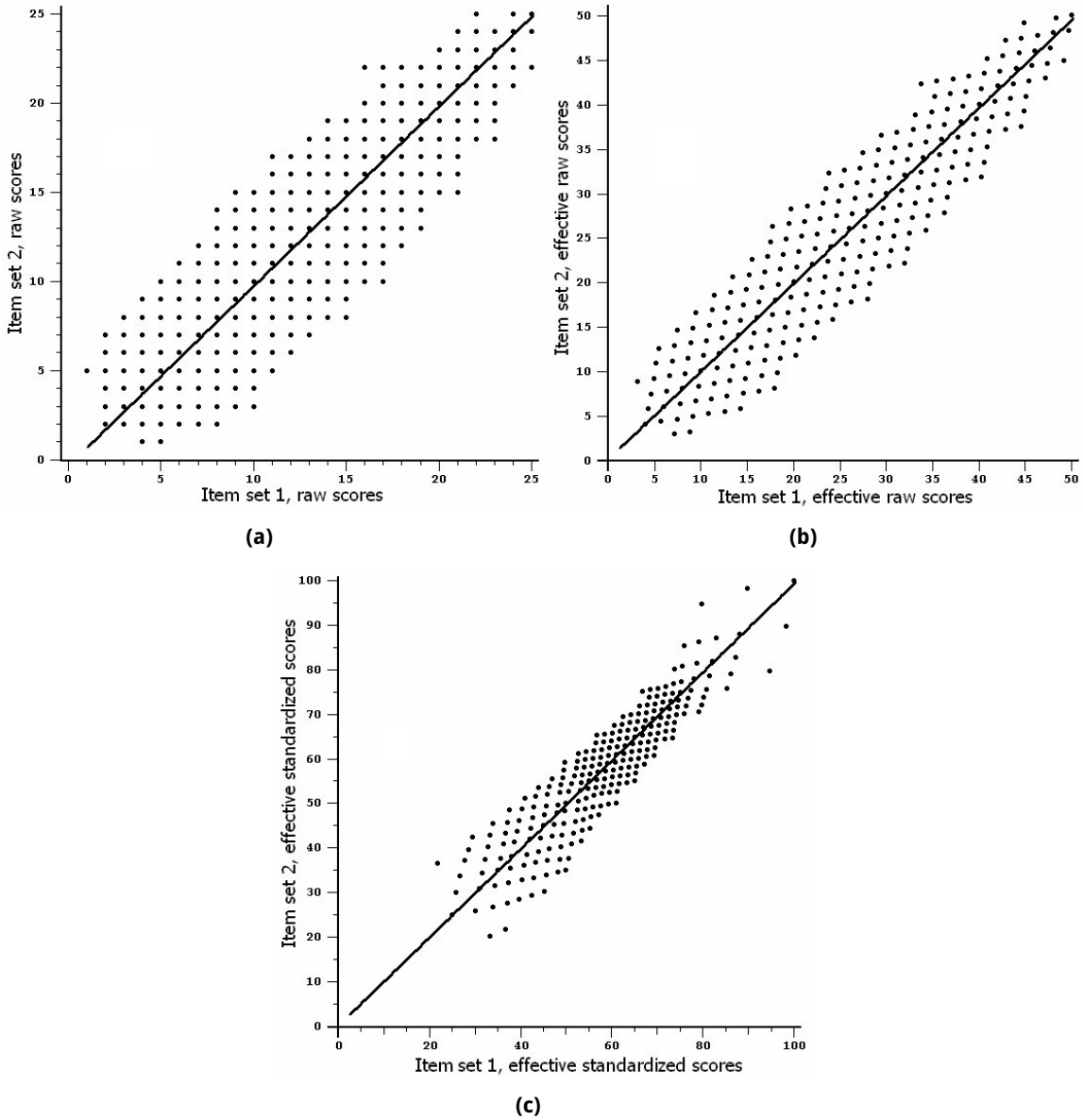
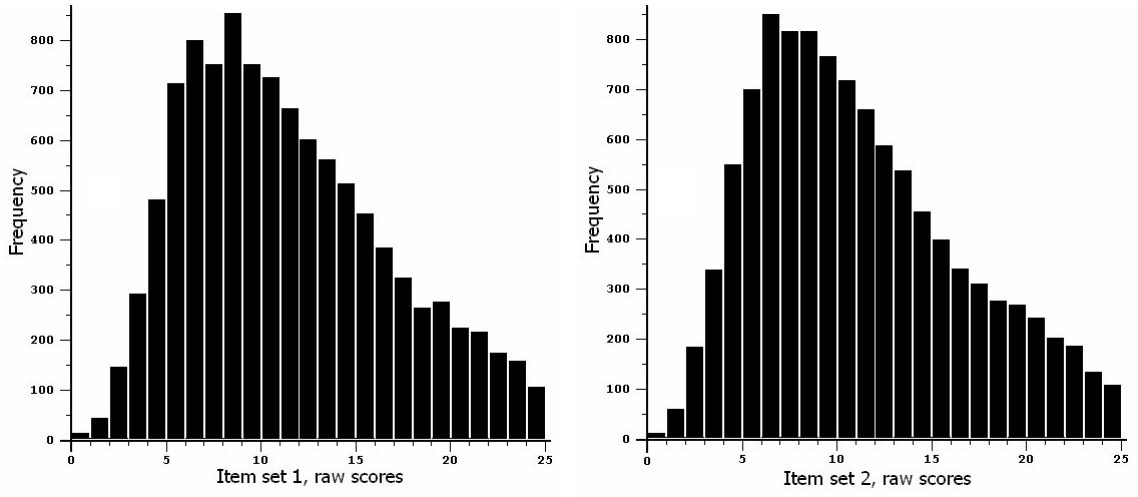


Figure 9. Relationship between scores for set 1 and scores for set 2: a) half-test raw scores; b) effective full-test raw scores; c) effective full-test scaled scores

Figure 10 represents a histogram showing the test score frequencies, i.e. number of participants who have obtained specified score for each item set (prior to data manipulation, correspond to figure 9a).

Comparison of the histograms in Figure 10 and 8 provides evidence of the proposed approach to comprising parallel test versions being successful.



(a) (b)

Figure 10. Test score frequencies for half-test item set 1 (a) and item set 2 (b)

This is confirmed by very similar percentage of correct answers: 43.9 % for set 1 and 42.8 % for set 2. Linear regression parameters confirm this conclusion too:

$$a = 1.00,\ b = -0.32,\ R = 0.863,\ w_x = 2.92,\ w_y = 2.92,\ \sigma = 2.06.$$

After conversion of raw scores into effective full-test raw scores in the 0-50 range:

$$a = 1.00,\ b = -0.45,\ R = 0.925,\ w_x = 4.13,\ w_y = 4.13,\ \sigma = 2.9.$$

Growth of the correlation coefficient is due to decrease by $\sqrt{2}$ times the distance between each of the point and the regression line after scaling using formulae (9), (10). After conversion to scaled test scores:

$$a = 1.00,\ b = -0.74,\ R = 0.911,\ w_x = 5.53,\ w_y = 5.53,\ \sigma = 3.9.$$

Average standard error is $\sigma = 3.9$ test scores. Thorndike's approach provided the same error value. When using odd and even item sets the value was 2.5 % higher: $\sigma = 4.0$ test scores. Thus, despite the fact that odd and even item sets were not ideally parallel the difference in the error estimate can be considered very small.

10. ERROR OF MEASUREMENT OF SCALED SCORES FOR DIFFERENT SCORE RANGES

For points of item set 1 and set 2 in the range from 53 to 75 scaled test scores (where the conversion is linear) the standard deviation was $w = 4.81$ scores which is less than for other ranges. Hence, the standard error for this range is $\sigma = 3.4$ scores, which is different from the value obtained by using odd and even item sets ($\sigma = 3.35$) by only 1.5 %.

Outside the linear range (below 53 and above 75 scores) the dispersion increases (Fig. 9c) in the same way as in Figure 3. The error of measurement increases accordingly. For the 0–53

test score range $w = 5.91$ scores and standard error of measurement is $\sigma = 4.18$ scores. This is different from the previously obtained value by only 1.4 %: by using odd and even item sets $\sigma = 4.24$ scores. Based on these values we can assume that $\sigma = 4.2$ scores for this range. For the 75-100 score range w= 7.1 scores and standard error is $\sigma = 5.0$ scores; the value obtained using odd and even item sets was $\sigma = 5.6$ scores which is different by 11 %. Therefore we can assume that for this range the error is $\sigma = 5.3$ scores (we took the average value for the two splitting methods).

Thus, different approaches to splitting the entire test into parallel test forms provide almost identical estimates of measurement error. Error of this estimate does not exceed 11 % and in most cases is less than 3 %.

11. SUMMARY

A method for estimating the error of measurement based on special splitting the entire test into two parallel forms and the subsequent conversion of half-test results is proposed.

A method of data processing and visualization based on the removal of the data corresponding to the most infrequent values is proposed. In the case of very noisy data this method allows eliminating the contribution of atypical values and provides a significant increase in the visibility of scatter plots.

References

1. L. S. Feldt, M. Steffen, and N. C. Gupta, "A comparison of five models for estimating the standard error of measurement at specific score levels," *Applied Psychological Measurement*, vol. 9, no. 4, pp. 351–361, 1985; doi: 10.1177/014662168500900402
2. F. M. Lord, "Estimating test reliability," *Educational and Psychological Measurement*, vol. 15, no. 4, pp. 325–336, 1955; doi: 10.1177/001316445501500401
3. J. A. Keats, "Estimation of error variances of test scores," *Psychometrika*, vol. 22, no. 1, pp. 29–41, 1957; doi: 10.1007/BF02289207
4. J. Ludbrook, "Comparing methods of measurement," *Clinical and Experimental Pharmacology and Physiology*, vol. 24, no. 2, pp. 193–203, 1997; doi: 10.1111/j.1440-1681.1997.tb01807.x
5. V. V. Monakhov, A. V. Kozhedub, P. A. Naumenko, L. A. Evstigneev, M. A. Krukelis, D. V. Solodovnikov, and I. B. Kernitskii, "BARSIC: A Programming System for Physicists," *Programming and Computer Software*, vol. 31, no. 3, pp. 157–165, 2005; doi: 10.1007/s11086-005-0028-2
6. V. V. Monakhov, S. K. Stafeev, L. A. Evstigneev, A. F. Kavtrev, and V. E. Fradkin, "The purpose and experience of online contests in physics," *Physics in Higher Education*, no. 4, pp. 53–63, 2007 (in Russian).
7. V. V. Monakhov, "Analysis of the results of Russian Unified State Examinations in mathematics and physics and online competitions in Physics," *Computer Tools in Education*, no. 1, pp. 50–57, 2011 (in Russian).
8. A. L. Qualls-Payne, "A Comparison of Score Level Estimates of the Standard Error of Measurement," *Journal of Educational Measurement*, vol. 29, no. 3, pp. 213–225, 1992; doi: 10.1111/j.1745-3984.1992.tb00374.x
9. G. Rasch, "On General Laws and the Meaning of Measurement in Psychology," In *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV*, Berkeley, CA, 1961, pp. 321–333.
10. R. L. Thorndike, "Reliability," In *Educational measurement*, E. F. Lindquist Ed., Washington DC: American Council on Education, 1951, pp. 560–620.
11. B. D. Wright and M. H. Stone, *Measurement Essentials*, 2nd Ed., Wilmington, DE: Wide Range, 1999.
12. D. Woodruff, "Conditional Standard Error of Measurement in Prediction," *Journal of Educational Measurement*, vol. 27, no. 3, pp. 191–208, 1990; doi: 10.1111/j.1745-3984.1990.tb00743.x

Received 25.07.2018, the final version — 06.09.2018.

Компьютерные инструменты в образовании, 2018
№ 5: 24–40
УДК: 519.25 : (378.146+004.4)
http://ipo.spb.ru/journal
doi:10.32603/2071-2340-2018-5-24-40

ОБРАБОТКА И ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ ТЕСТОВ

Монахов В. В.[1], Кожедуб А. В.[1], Ханнанов Н. К.[2], Королёв А. А.[3], Курашова С. А.[3]

[1]Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
[2]Автономная некоммерческая общеобразовательная организация «Новая Черноголовская школа», Черноголовка, Россия
[3]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия

Аннотация

Предложен метод обработки данных для оценки погрешности измерения шкалированных (тестовых) баллов. Он включает в себя разделение заданий на два параллельных варианта (полутеста), масштабирование первичных баллов полутестов до эффективных первичных баллов по полному тесту и преобразование их в шкалированные тестовые баллы. Показано, что метод позволяет с высокой точностью оценить погрешность измерения. Предложен подход к визуализации данных, который использует удаление части данных, соответствующих наиболее редким значениям. В случае сильно зашумленных данных этот метод помогает устранить вклад нетипичных значений и обеспечивает значительное увеличение наглядности графиков рассеяния.

Ключевые слова: *обработка данных, визуализация данных, компьютерное тестирование.*

Цитирование: Монахов В. В., Кожедуб А. В., Ханнанов Н. К., Королёв А. А., Курашова С. А. Обработка и визуализация результатов тестов // Компьютерные инструменты в образовании, 2018. № 5. С. 24–40 .

Поступила в редакцию 25.07.2018, окончательный вариант — 06.09.2018.

Монахов Вадим Валериевич, доцент Санкт-Петербургского государственного университета, физический факультет, кафедра вычислительной физики; 198504, Россия, Санкт-Петербург, ул. Ульяновская, 1, НИИ физики, v.v.monahov@spbu.ru
Кожедуб Алексей Владимирович, старший преподаватель Санкт-Петербургского государственного университета, физический факультет, кафедра вычислительной физики; 198504, Россия, Санкт-Петербург, ул. Ульяновская, 1, НИИ физики, a.kojedub@spbu.ru
Ханнанов Наиль Кутдусович, заместитель директора по воспитательной работе, Автономная некоммерческая общеобразовательная организация «Новая Черноголовская школа»; 142432, Россия, Черноголовка, Спортивный бульвар, 9, newschoolchg2017@gmail.com

Королёв Александр Александрович, доцент, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; 197101, Россия, Санкт-Петербург, Кронверский проспект, 49, korolev3010@mail.ru
Курашова Светлана Александровна , старший преподаватель, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; 197101, Россия, Санкт-Петербург, Кронверский проспект, 49, sakurashova@yandex.ru

Vadim V. Monakhov,
Associate Professor, Saint Petersburg State University, Department of Computational Physics, Institute of Physics; 198504, Russia, Saint Petersburg, Ulyanovskaya, 1,
v.v.monahov@spbu.ru

Alexey V. Kozhedub,
Senior Lecturer, Saint Petersburg State University, Department of Computational Physics, Institute of Physics; 198504, Russia, Saint Petersburg, Ulyanovskaya, 1,
a.kojedub@spbu.ru

Nail K. Khannanov,
Deputy Director for educational work, Autonomous nonprofit general education organization "New Chernogolovskaya School"; 142432, Russia, Chernogolovka, Sportivny Boulevard, 9,
newschoolchg2017@gmail.com

Alexander A. Korolev,
Associate Professor, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics; 197101, Russia, Saint Petersburg, Kronverkskiy prospekt, 49,
korolev3010@mail.ru

Svetlana A. Kurashova,
Senior Lecturer, Saint Petersburg National Research University of Information Technologies, MechanicsA and Optics; 197101, Russia, Saint Petersburg, Kronverkskiy prospekt, 49,
sakurashova@yandex.ru