A Method for the Age Standardization of Test Scores

I. P. Schagen
National Foundation for Educational Research, England

A procedure is presented to generate standardized scores from raw test data that are, as far as possible, age independent and normally distributed. The model is fitted to the percentile points of the raw score distribution, and assumes a linear trend of each percentile with age. The fitted slopes can be constant or can vary quadratically with the percentiles. A nonlinear transformation of the data is also possible to allow for "ceiling effects." These models are described and the methods used to fit them to test data are discussed; examples are presented of their use in standardizing tests, and the use of the diagnostic plots produced by the program are discussed. *Index terms: age standardization, linear regression, nonlinear regression, nonparallel regression, parallel linear regression, percentiles, score transformation.*

Testing of both children and adults is currently a growth industry. In many cases, tests are administered to individuals who differ significantly in age, and some allowance must be made for this when norms are developed. This involves fitting a suitable model to the test scores, which should be both sophisticated enough to fit the data adequately yet simple enough to be acceptable to users of the test.

The basic model discussed here involves a linear trend for each percentile point of the score distribution with age. The simplest version assumes that the slopes are identical for all percentiles, but in some cases this model is not adequate. An extension of the basic model assumes differential slopes, and may provide an accurate representation of the data. Alternatively, nonlinear (log-odds) transformations of the data can be applied to model "ceiling effects." From the fitted model of whatever type, it is possible to generate standardized scores that are assumed to be age independent and normally distributed.

The Standardization Model

The basic concept of the model is to use the percentile points of the raw score distribution for each age group, and to fit a linear model of their variation with age. Healy, Rasbash, and Yang (1988) developed a similar but more complex model for age-related data. Their model used smoothed estimates of the percentile points, whereas values estimated from discrete age groups were used here. The latter method is computationally simpler, and more acceptable in test standardization. The present models were also initially confined to linear functions.

To fit the model, a set of age groups must be defined, either in units of one month or any integer multiple, and a set of score groups is needed that spans the range of possible scores.

Then

n_{ij} = number of examinees in the ith age group and the jth score group,

$$N = \sum_i \sum_j n_{ij} \quad , \tag{1}$$

© Copyright 1990 Applied Psychological Measurement Inc.
0146-6216/90/040387-07$1.60

$$n_i = \sum_j n_{ij} \quad , \tag{2}$$

a_i = midpoint of the ith age group,
s_j = upper limit of the jth score group, and
n_s = the number of score groups.

Define a set of quantile values, x_k, $k = 1, \ldots, m$ [possibly such that $x_k = k/(m + 1)$]. For the ith age group, compute the percentile points Q_{ki}, $k = 1, \ldots, m$ as follows:

$$\text{Form } f_1 = n_{i1}/n_i \quad , \tag{3}$$

$$f_j = f_{j-1} + n_{ij}/n_{i.} \quad , j = 2, n_s \quad , \tag{4}$$

Interpolate the table (f_j, s_j) with the value x_k to get the percentile point Q_{ki}.

The model to be fitted is:

$$Q_{ki} = \bar{Q}_k + b(a_i - \bar{a}) + \epsilon_{ki} \quad , \tag{5}$$

where $\bar{Q}_k$ is the percentile point at the mean age $\bar{a}$, b is the age allowance, and ϵ_{ki} is a residual error with mean 0.

To fit this model, parallel linear regression can be used, with weights proportional to the numbers in the age groups—that is

$$w_i = n_i/N \quad . \tag{6}$$

Then

$$\bar{Q}_k = \sum_i w_i Q_{ki} \quad , \tag{7}$$

$$\bar{a} = \sum_i w_i a_i \quad , \tag{8}$$

$$S_{aa} = m \sum_i w_i (a_i - \bar{a})^2 \quad , \tag{9}$$

$$S_{qa} = \sum_k \sum_i w_i (Q_{ki} - \bar{Q}_k)(a_i - \bar{a}) \quad , \text{ and} \tag{10}$$

$$b = S_{qa}/S_{aa} \quad . \tag{11}$$

To obtain standardized scores, let Z_k be the standard-normal value equivalent to x_k, and μ and σ be the required mean and standard deviation.

Then

$$Y_k = \mu + \sigma Z_k \quad . \tag{12}$$

For a given raw score s and age a, the procedure to obtain a standardized score y is to compute the value

$$s^* = s - b(a - \bar{a}) \quad , \tag{13}$$

and then to interpolate the table $(\bar{Q}_k, Y_k)$ with s^* to get y. (Note that standardized scores are normally computed as rounded to the nearest integer.)

Developments From the Basic Model

The basic age standardization model can be amended in several ways to deal with particular situations. One situation is the question of weighting. Each age group is weighted by the number of examinees it contains, but by default each percentile is given the same weight in computing the overall slope of the regression line. However, percentiles close to the center of the range are more accurately defined than those at the extremes; thus it would seem reasonable to allow for this fact. Therefore, the model allows for the option of weights to be applied to the percentiles in the regression only. These weights are calculated to reflect the closeness of the percentile value to .5.

Normally, it is assumed that each examinee is given the same weight (i.e., is equally representative of the population). However, in some circumstances this may not be true, and examinees may have to be assigned individual weights. These weights may be input with the test data and applied in calculating the percentile points and the weights for each age group. Thus there may generally be different sets of weights for age groups, for percentiles, and for examinees.

The fundamental assumption of the basic standardization model is that the percentile values may be fitted by a set of parallel straight lines as a function of age. There may be circumstances, however, in which this assumption is not valid. For example, if there is a "ceiling effect," then the test is too easy for the age range and the top percentiles are unable to increase with age at the same rate as the lower percentiles. Various solutions are possible to this type of problem, but the simplest is to fit nonparallel straight lines to the data. An alternative solution is to use a transformation of the data (e.g., a log-odds transformation).

Nonparallel Regression Model

It is possible to estimate the slope for each percentile separately, but it is not necessarily a good idea to use these estimated values directly. Extreme percentiles generally have higher standard errors in their slopes, and it is better to use a consistent model to smooth the slope estimates. The method selected was to fit a quadratic model, linking slope of regression line to score at mean age for each percentile value. Figure 1 shows some sample data of this type, and indicates that a quadratic model gives a reasonable fit. Both high and low scoring examinees tend to have lower age allowances than those in the middle. A simple calculation allows conversion of raw score and age to score at mean age and thus to standardized score.

Let

s^* = score at mean age,

s = raw score at age a,

$$x = \bar{a} - a \quad , \tag{14}$$

b = slope of the regression line

$$= \alpha_0 + \alpha_1 s^* + \alpha_2 (s^*)^2 \quad . \tag{15}$$

Then

$$s^* = s + bx$$

$$= s + x[\alpha_0 + \alpha_1 s^* + \alpha_2 (s^*)^2] \quad . \tag{16}$$

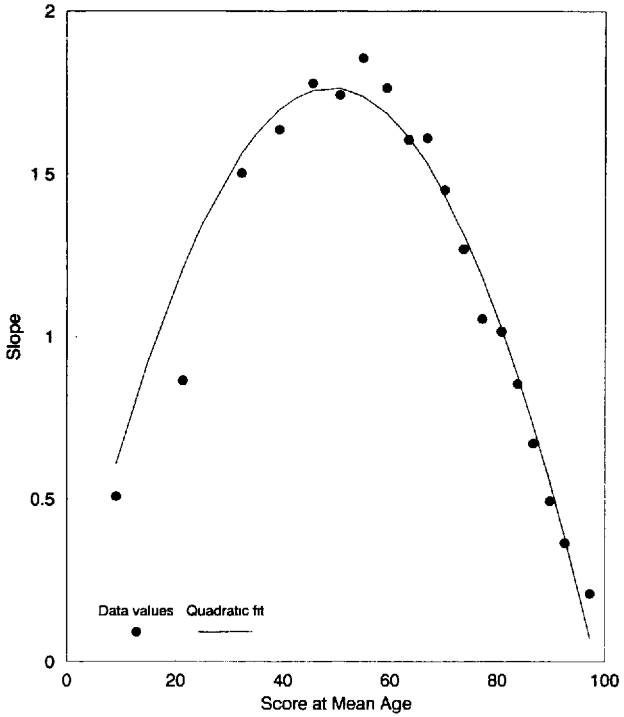
Therefore,

$$\alpha_2 x (s^*)^2 + (\alpha_1 x - 1) s^* + \alpha_0 x + s = 0 \quad , \tag{17}$$

and the solution is:

$$s^* = \frac{1 - \alpha_1 x - [(\alpha_1 x - 1)^2 - 4\alpha_2 x(\alpha_0 x + s)]^{1/2}}{2\alpha_2 x} \quad . \tag{18}$$

Figure 1
Regression Slope as a Function of Score



Nonlinear Transformation

The advantage of linear regression in these applications is its simplicity, together with an easy interpretation of the regression slope in terms of "age allowance." However, a principal drawback is the fact that a linear increase of test score with age is only a reasonable assumption over a limited range of ages, the range being determined by the individual test. The worst problem occurs when the ceiling effect occurs, because linear models will then predict scores higher than the maximum. The nonparallel regression model attempts to circumvent this problem, but is again successful only over a limited range. A point is reached at which percentile lines will clearly start to cross; this means that the model has broken down totally.

One alternative is to abandon the simplicity of linear regression in favor of another model. Obviously, quadratic or other polynomial regression is not a reasonable solution, because such curves would result in problems similar to those of nonparallel regression. In addition, there is no guarantee that the regression functions will be nondecreasing with age.

A better nonlinear, nondecreasing function of age is obtained by transforming the data using a "log-odds" transform. For any raw score s, with a maximum score of S on the test, define the transformed value

$$t = -\ln[(S - s)/s] \quad , \tag{19}$$

which takes on values from $-\infty$ to ∞ as s goes from 0 to S. Fitting a linear function to the transformed t values is equivalent to a nonlinear but increasing function of age for the original scores.

However, the percentile lines fitted in this way, if parallel in *t*, cannot cross. Conversely, these lines can never rise above the maximum score *S*; thus, the ceiling effect problem is dealt with.

In principle there are now four models possible to fit to test data:

1. Untransformed, parallel.
2. Untransformed, nonparallel.
3. Transformed, parallel.
4. Transformed, nonparallel.

In practice the fourth model is never used because Models 2 and 3 are alternative ways of dealing with ceiling-effect problems; if neither model is effective, the data must be treated with extreme caution.

For well-designed tests over a limited age range, the first model is normally adequate, provided no extrapolation to ages outside those tested is required, and no ceiling effect occurs. Otherwise, either Models 2 or 3 may be fitted. Model 2 is used with limited-range data and moderate ceiling effect if the advantages of using untransformed data overcome the disadvantages of the linear model. The third model is generally the most robust and applicable, provided the use of nonlinear "age allowances" is acceptable.

Examples of Model Application

The program STANEW has been written in FORTRAN to implement this approach to the age standardization of tests. It can implement all stages of the process, from the input of the basic data, through scoring the test, forming age groups, estimating percentiles and fitting regression lines with

Figure 2
Parallel Linear Model Fitted to Suffolk Reading Scale Data

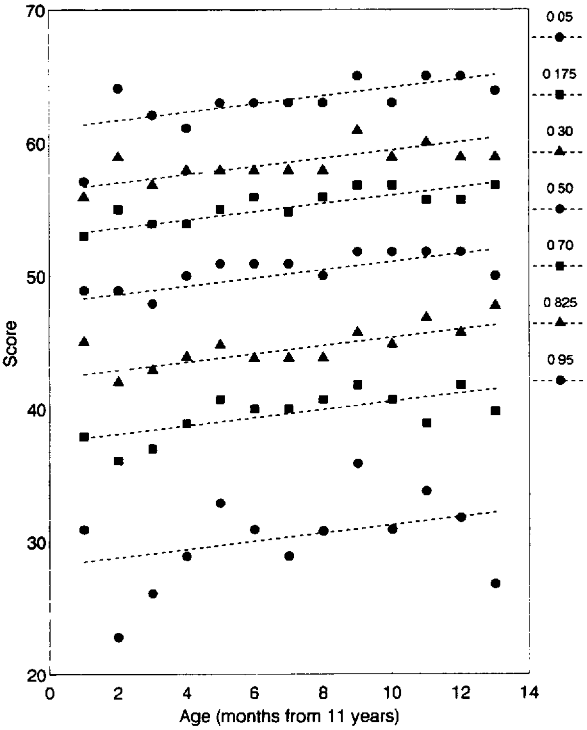
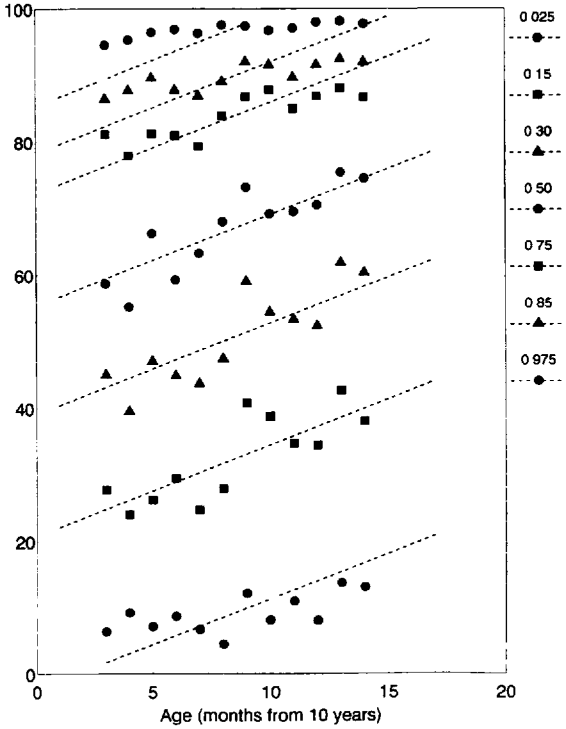


Figure 3
Parallel Linear Model Fitted to Devon Test 7 Data



age, to the output of standardized scores and tables for manual conversion. During this process a "diagnostic plot" is output as a means of indicating the applicability of the model to the current dataset.

The first example to show the application of the program to real test data makes use of results from the Suffolk Reading Scale Autumn standardization, based on scores from 1,852 examinees. The Suffolk Reading Scale is a multiple-choice cloze test that presents single sentences and five options to complete the blanks (Hagley, 1987). The diagnostic plot for this dataset is shown in Figure 2, and consists of percentile values for the different age groups, plus fitted regression lines.

The program has been used on a wide variety of datasets, and in most cases the results obtained with the simple parallel line model have been entirely acceptable. A few exceptions have occurred in which the assumption of parallel trends against age for each percentile has been refuted. One example is shown in Figure 3, in which a parallel linear model has been fitted. This dataset is derived from the Devon Verbal Reasoning Test 7 results for 1,996 examinees (used to determine entry to selective secondary schools), and is the same one used to produce Figure 1. The results in Figure 3 illustrate a quadratic relationship between age trend and score at mean age. In Figure 4, the diagnostic plot for the nonparallel linear model is shown, and it appears to give an acceptable fit to the data. For comparison, Figure 5 shows the plot for the model fitting to transformed data; this gives an acceptable fit to the data over the available age range, and extrapolates to other ages in an appropriate fashion. Caution should obviously be used, however, in using such models for ages over which they have not been fitted.

Figure 4
Nonparallel Linear Model Fitted to Devon Test 7 Data

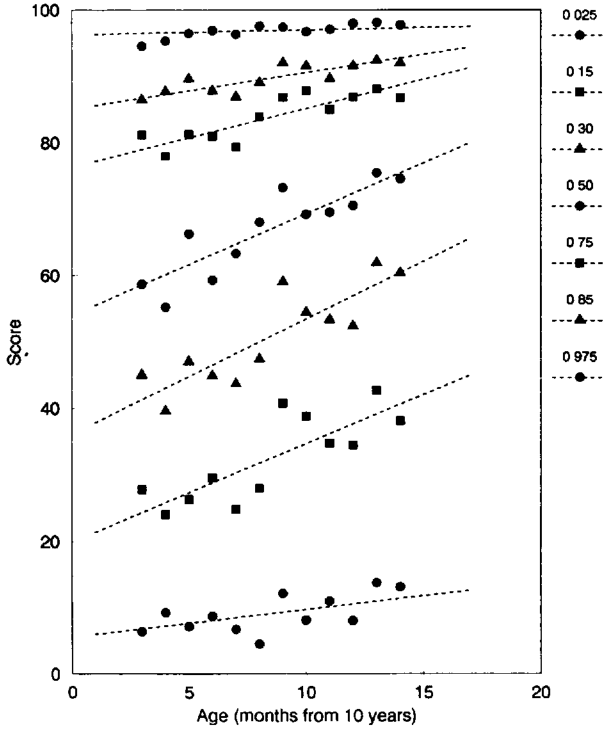
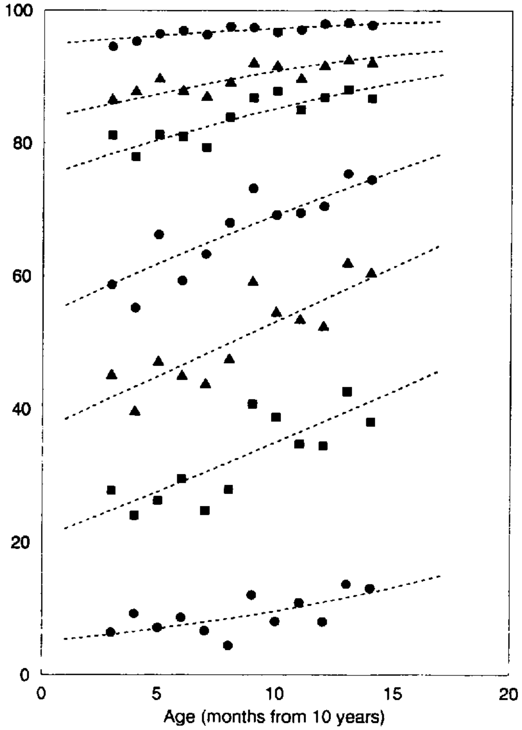


Figure 5
Nonlinear Model Fitted to Devon Test 7 Data



Experience with the program STANEW has shown that it is able to fit models to a wide variety of test data over age ranges of three years or more. The program was developed initially because acceptable alternatives did not exist. The only method previously available was devised by Lawley (1950) as a hand-calculation approach. It fits a simple linear age allowance to a single percentile value, and then transforms the whole dataset to approximate normality. There are a number of problems with this approach; for example, it takes no account of ceiling effects or variations in distribution with age, and it does not allow the assumptions of the model to be reviewed critically by means of a diagnostic plot.

Conclusions

The method described above for the age standardization of test scores appears to give reasonable and consistent results when applied to a range of different datasets. Obviously, the method should not be used with data that violate the basic assumptions on which any attempt at age standardization must rest. For example, if examinees come from two or more disparate populations with different age distributions, then the results of this model may well be quite misleading.

The simple linear age-allowance model appears to be reasonable over a wide age range—sometimes as much as four years. However, this may not be the case for all types of tests. Nonlinear transformations can be used to deal with departures from the simple model, particularly ceiling effects. The diagnostic plots provided by program STANEW should be examined, and the model should then be modified, if necessary.

References

Hagley, F. (1987). *Suffolk Reading Scale: Teacher's guide.* Windsor: NFER-Nelson.

Healy, M. J. R., Rasbash, J., & Yang, M. (1988). Distribution-free estimation of age-related centiles. *Annals of Human Biology, 15,* No.1, 17–22.

Lawley, D. N. (1950). A method of standardizing group tests. *British Journal of Psychology (Statistical section), 3,* 86–89.

Author's Address

Send requests for reprints or further information to Ian P. Schagen, National Foundation for Education Research, The Mere, Upton Park, Slough Berkshire SL1 2DQ, England.