Chapter Four

Scoring the TAGG

Recommended APA-Style Citation

Martin, J., Hennessey, M., McConnell, A., Terry, R., & Willis, D. (2015). *TAGG technical manual*. Retrieved from https://tagg.ou.edu/tagg/

Chapter Four

Scoring the TAGG

The purpose of this chapter of the Technical manual is to present the steps employed to develop scores for the three versions (i.e., professional, family, and student) of the *Transition Assessment and Goal Generator* (TAGG). Procedures were the same for the three versions. Due to their proprietary nature, specific scoring algorithms will not be presented. Rather, an overview of the approaches used to develop scores and generate results for the three versions of the TAGG will be presented here. This chapter will be organized in the following way

- Description of the specific challenges associated with scoring the TAGG and the decisions made to alleviate these challenges;
- A description of the four-step procedure employed to create each of the TAGG construct (subscale) scores and compare the scale scores across constructs; and,
- Descriptions of the procedures used to generate scoring profile results, including specific goals for each student based on their scale scores.
- A presentation of handling of missing data is also included.

All tables and figures are presented at the end of the chapter.

Decisions Made to Alleviate Scoring Challenges

Formulating scores of the TAGG-P, TAGG-F, and TAGG-S presents several psychometric challenges. Although the structures of the TAGG instruments are well-understood through factor-analytic study, scoring represents the psychometric goal of placing each student onto a common scale which can then be used for communicating actions to be taken. Among the many challenges present in the TAGG are a) the fact that each subscale of the TAGG contains a different number of items, and b) some items are scored on different scale types (e.g., 5-point Likert-type scales versus Yes/No binary items). Moreover, although the factor analytic results represent a good first-order approximation to the true structure, it is likely the item responses are actually related non-linearly to the latent constructs measured by the three versions of the TAGG.

To address these scaling issues with the TAGG (all three versions), we employed Item Response Theory (IRT) to score each subscale of the TAGG. IRT scaling has certain advantages to the classical rules of scoring tests by summing items. These advantages include the ability to scale different item types, provide a common

metric for scales with different numbers of items, weight items differentially by their validity for assessing the construct of interest, and obtain sample-invariant estimates of the item parameters used in creating scores.

In applying IRT principles to the TAGG, we made several choices among the various IRT technologies available in the literature. First and foremost, we chose to use Samejima's (1969) graded response model as the basis for estimating the item parameters. Samejima's model handles both polytomous response and binary response data with a minimum of scaling assumptions, making it a natural choice for the TAGG.

Second, given the item parameters, we chose to create IRT scale scores from summed scores using the approach of Thissen, Pommerich, Billeaud, and Williams (1995). Typical IRT-scoring uses response-pattern scoring, where unique response patterns result in a unique IRT scale score. Because of the need to create scores quickly in real-time assessment, estimates of construct-level scores are instead obtained via summed score approximations. These so-called EAP|SS (expected a posteriori scores for a given summed score) scores are basically weighted averages of the full EAP scores for all response patterns that result in the same summed score. Although some precision will be lost, considering only 25 (5 x 5) possible summed scores versus 3,125 (5^5) possible response patterns in a 5-item 5-point Likert-type response scale is crucial when implementing real-time scoring.

Finally, we chose to use IRTPRO software (Cai, Thissen, & du Toit, 2011) to construct the scoring tables needed for each subscale construct for each version of the TAGG. These tables were then used as the basis of the algorithms implemented via the web-based TAGG software to create scoring displays and to assist in creating the appropriate transition goals for the student.

Creating the TAGG Subscale Scores

Before describing the procedures employed to develop TAGG subscale scores, it is appropriate to present a description of the sample of participants used to develop these procedures. Data from all participants responding to the three versions of the TAGG in Phases I and II of data collection were included in the present analysis. Because further details regarding the characteristics of the sample are presented in other chapters in the Technical Manual, only basic demographic characteristics will be presented here. Readers are referred to Chapters 3 and 5 for more sample details.

Professional Participants

Thirty-nine transition professionals participated in data collection in Phase I and an additional 34 participated in Phase II. Of these transition professionals, a total of 68 reported being female. Additional sample characteristics can be found in Table 1.

Family Participants

Results presented in this chapter of the Technical Manual relating to the development of scores and scoring procedures are based on a total of 500 family member participants. Specifically, 271 family members participated in Phase I and 229 participated in Phase II of this research. Further sample characteristics are presented in Table 2.

Student Participants

The algorithms and procedures presented here are based on data collected from 691 student participants. Three hundred forty nine participants were included in Phase I and the remaining 342 students participated in Phase II. More demographic information

is included in Table 3 and in Chapters 3 and 5 of the Technical Manual.

Description of the Algorithms and Procedures Employed

Using the IRT methodology described previously, we constructed a four-step algorithm for converting raw data into scores for each student to assist in setting transition goals. These four steps include 1) placing each scale onto a common score metric, 2) projecting item characteristics (e.g., item difficulty) onto the scale score metric, 3) conducting a within-student comparison of scale scores across constructs to determine relative strengths and weaknesses, and 4) conducting a within-construct comparison of a student's scale score to item responses (e.g., difficulty) to generate appropriate goals on identified weaknesses. We will now consider each of these steps in more detail. To illustrate these steps, we will use the TAGG-P Disability Awareness (DA) scale throughout this section.

Placing Each Scale Onto a Common Metric

When identifying relative strengths and weaknesses across constructs, it is important to compare scores in the same scale. As noted earlier, we used the Thissen et al. (1995) EAP|SS methodology to create scale scores for each construct, even though each construct had a different number of items and, thus, a different summed-score range. The result is a standard z-score scale for each construct. Figure 1 represents a graphical depiction of the process using the IRT-based Test Characteristic Curve (TCC), which shows the (non-linear) relationship between summed score and scale score.

Table 4 shows the specific results of scoring the TAGG-P DA scale using the EAP|SS transformation. Reading Table 4, a raw summed score of 0 (on a 0-4 scale) results in a scale score of $\Theta = -2.012$, indicating poor responding in terms of Disability Awareness. Conversely, a raw summed score of 16 results in a scale score of $\Theta = +1.862$, indicating a strong positive response on this scale. The standard error of measurement for each scale score is used later in determining relative strengths and weaknesses.

For each construct on each of the three TAGG versions, we implemented this process, creating tables of scale scores and implementing these tables in the web-based version of the TAGG-P, TAGG-F, and TAGG-S. These scale scores will then be compared within a student's results to identify relative strengths and weaknesses.

Projecting Item Characteristics Onto the Scale Score Metric

One of the advantages of using IRT methodology is the capability of comparing each student's scale score to the relative response propensity (e.g., item difficulty) of each item. This capability allows the TAGG to identify specific behaviors for which a student shows a relative weakness once the overall behavioral construct has been identified as a weakness. Technically, of course, these comparisons are best made when a Rasch (equal-slope) version of the Samejima model holds. For the TAGG, item slopes did vary across items; nevertheless, a first-order approximation (an average) of item difficulty can be constructed and used for the purpose of identifying shortcomings in specific behavior. Figure 2 shows a graphical description of the item characteristic curves for the four items of the TAGG-P DA scale, and Figure 3 shows the averaged item difficulty for item 2 of the TAGG-P DA Scale. The arrow on Figure 3 shows the value on the proficiency scale (e.g. scale

score) for which a student is at least 50% likely to respond with a 2 or higher on a 0-4 scale. We have called this scale-score value the average item difficulty for the item. This suggests a student with a scale score greater than -0.24 will be more likely than not to respond in the upper half of the scale. This methodology was used to create average item difficulties for each item and the results are stored in a table in the TAGG web-based software.

Comparing Scale Scores Across Constructs

As stated earlier, we now want to conduct a *within-student comparison* of scale scores across constructs to determine relative strengths and weaknesses. Since each of our behavioral constructs are now represented on a common metric, we can simply rank order each student's scores *across the constructs* and use the lowest score to determine each student's relative weaknesses and relative strengths. This section encompasses steps 3 and 4 of the four-step procedure employed to develop TAGG scores. Figure 4 shows a visual representation of this process comparing only two constructs for ease of presentation.

Figure 4 shows an example of both a relative strength and a relative weakness on two constructs. In the figure, a student scores low (Θ = -1.10) on Disability Awareness while scoring high (Θ = +0.97) on Persistence. In this simple case, this student would be identified as having a relative weakness on Disability Awareness and having a relative strength on Persistence. Of course, on the TAGG-P, the software ranks up to eight different behavioral constructs before identifying relative strengths and weaknesses.

Finally, the TAGG software also considers the fact that although students will vary on their scale scores across constructs, they might not vary all that much. When students' scale scores vary by more than a standard error of measurement, a relative strength or weakness is labeled then as a greatest strength or weakness, rather than just a relative strength or weakness. Figure 4 contains two scale scores that are more than a standard error of measurement apart, and thus would be labeled as a greatest weakness (strength).

Generating Scoring Profile Results
Generating Specific Behavioral Goals

The last of the four-step algorithm involves conducting a *within-construct comparison* of a student's scale score to the average item responses (e.g., difficulty) to generate appropriate goals on identified weaknesses. Following step 3 of the process above, the TAGG software will have identified up to two relative (or greatest) weaknesses at the overall behavioral construct level. For goals to be of value, however, they must be specific. Because each of the items addresses specific behaviors, the TAGG compares the overall construct-level scale score (step 1) to the item-specific average difficulties. When a scale score fails to exceed an average item difficulty, that item is tagged to generate a specific behavioral goal addressing the content of that item. Of course, since there are many items per construct, there may be several instances in which a student's scale score fails to exceed the item-specific difficulty value. In these cases, we again rank order the differences between the student's scale score and all of the item-specific difficulty values, and then we choose those specific behaviors for which the student has indicated the poorest performance. This guarantees at least two goals are generated for each student-identified weakness.

Stanine Scores

Although all scoring and scoring comparisons are done at the scale-score level, presenting scale scores in a z-score metric is not advisable to the general public. For presentation purposes, all scale scores are transformed using a stanine transformation. The stanine transformation places scale scores into one of nine categories. Each of the nine categories has a width corresponding to a half of a standard deviation on the normal curve, with the mean lying at the center of the stanine score of 5. Stanine scores have the advantage of being single-digit scores and thus easy to graph, while reducing the tendency to try to interpret small scale-score differences. Figure 5 shows an example of how the stanine scores will be presented.

Missing Data

Although the TAGG software encourages each respondent to answer all questions, sometimes missing data occurs. While IRT pattern-scoring procedures do not require complete data to estimate individual scale scores, using the EAP|SS summed score estimates do require complete (or at least imputed) data for all items. To ascertain the effect of missing data on TAGG scores, a small simulation comparing the degree of missing data and various methods for imputation (within-student mean imputation, between-student mean imputation, regression-based estimates, etc.) was conducted on a set of compete TAGG data. Using a criterion of at least an $r = .90$ correlation between imputed and actual (complete) values, we determined that using a within-student (within-construct) mean imputation with no more than one missing value per construct returns reasonable estimates of scale scores. The web-based implementation of the TAGG uses this imputation algorithm when missing data occurs.

References

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows*. [Computer software]. Lincolnwood, IL: Scientific Software International.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (4, Pt. 2).

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

Table 1

Demographic Information for Professional Participants

	Phase I	Phase II
Sample size	39	34
Average Age	47 (10.2)	46 (8.7)
Average Years Teaching Experience	16 (10.9)	13 (9.2)
% Female	94.9	91.2
Racial/Ethnic Categories		
% Caucasian	76.9	85.3
% African American	12.8	11.8
% Hispanic	5.0	2.9
% American Indian	2.6	8.8

Note. Standard deviations are given in parentheses

Table 2

Demographic Information for Family Participants

	Phase I	Phase II
Sample Size	271	229
Average Age	45 (8.6)	43 (11.3)
Family Respondent		
% Mother/stepmother	80.0	78.0
% Father/stepfather	11.0	11.0
% Grandparent	3.7	5.7
% Legal guardian	2.6	0.9
% Lived with student	97.8	97.8
Family Education		
% No H.S. diploma	14.0	9.2
% H.S. education only	45.0	37.1
% Greater than H.S. education	38.0	51.5
% Had help with forms	0.9	6.6
Racial/Ethnic Categories		
% Caucasian	68.0	75.1
% African American	10.0	8.3
% Hispanic	6.0	5.2
% American Indian	3.0	10.9

Note. Standard deviations are given in parentheses

Table 3

Demographic Information for Student Participants

	Phase I	Phase II
Sample Size	349	342
Average Age	17 (1.4)	16 (3.1)
% Female	46.4	43.7
% Eligible for Free/Reduced Lunch	56.0	55.7
Grade Level		
% 9th grade	12.0	21.0
% 10th grade	26.0	16.3
% 11th grade	27.0	25.7
% 12th grade	35.0	35.6
Racial/Ethnic Categories		
% Caucasian	67.0	70.3
% African American	17.5	11.4
% Hispanic	12.0	11.4
% American Indian	4.0	13.4
% ELL	1.7	2.6
Disability Information		
% LD	61.0	56.6
% ID	12.0	13.1
% OHI	12.0	15.2
% ED	5.0	6.7
% Other disability	12.0	8.4
% Secondary disability	11.5	14.0

Note. Standard deviations are given in parentheses

Table 4

Scale Scoring Table for the Disability Awareness Subscale of the TAGG-P

Raw Score	Scale Score	Standard Error
0	-2.012	0.541
1	-1.589	0.458
2	-1.330	0.442
3	-1.104	0.431
4	-0.898	0.426
5	-0.699	0.420
6	-0.509	0.419
7	-0.323	0.422
8	-0.139	0.428
9	0.045	0.435
10	0.233	0.444
11	0.431	0.454
12	0.641	0.468
13	0.858	0.469
14	1.108	0.476
15	1.410	0.492
16	1.862	0.565

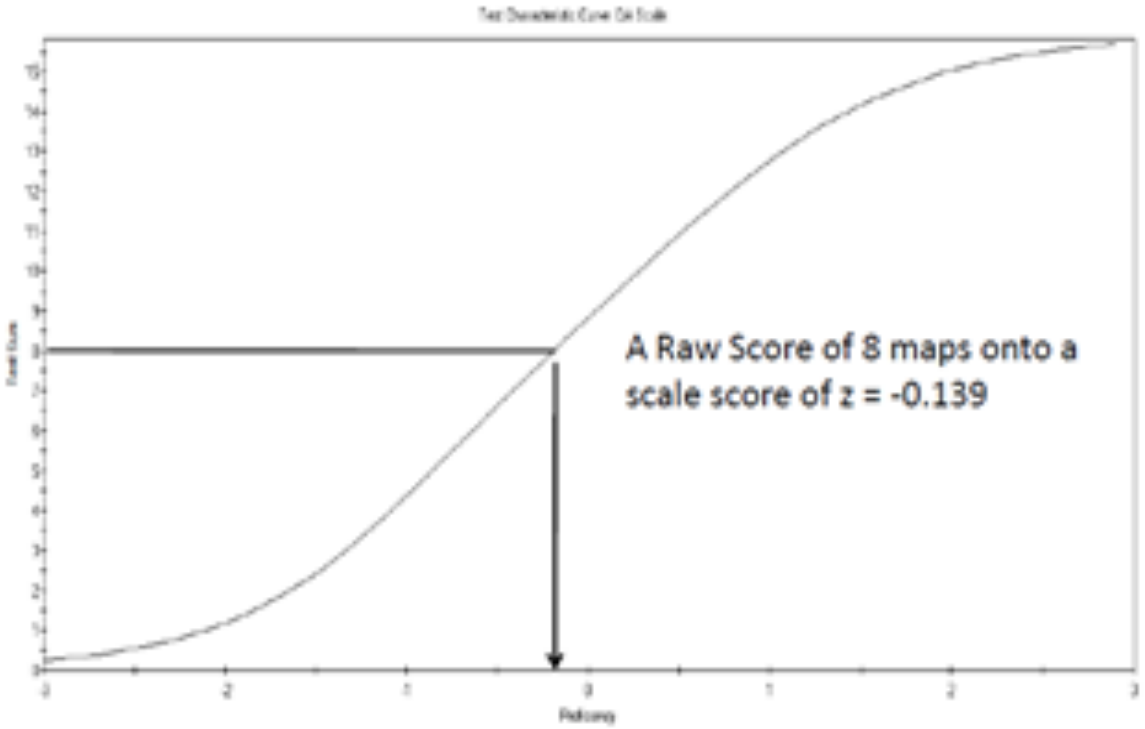

Figure 1. EAP|SS transformation via the TCC for TAGG-P DA scale.

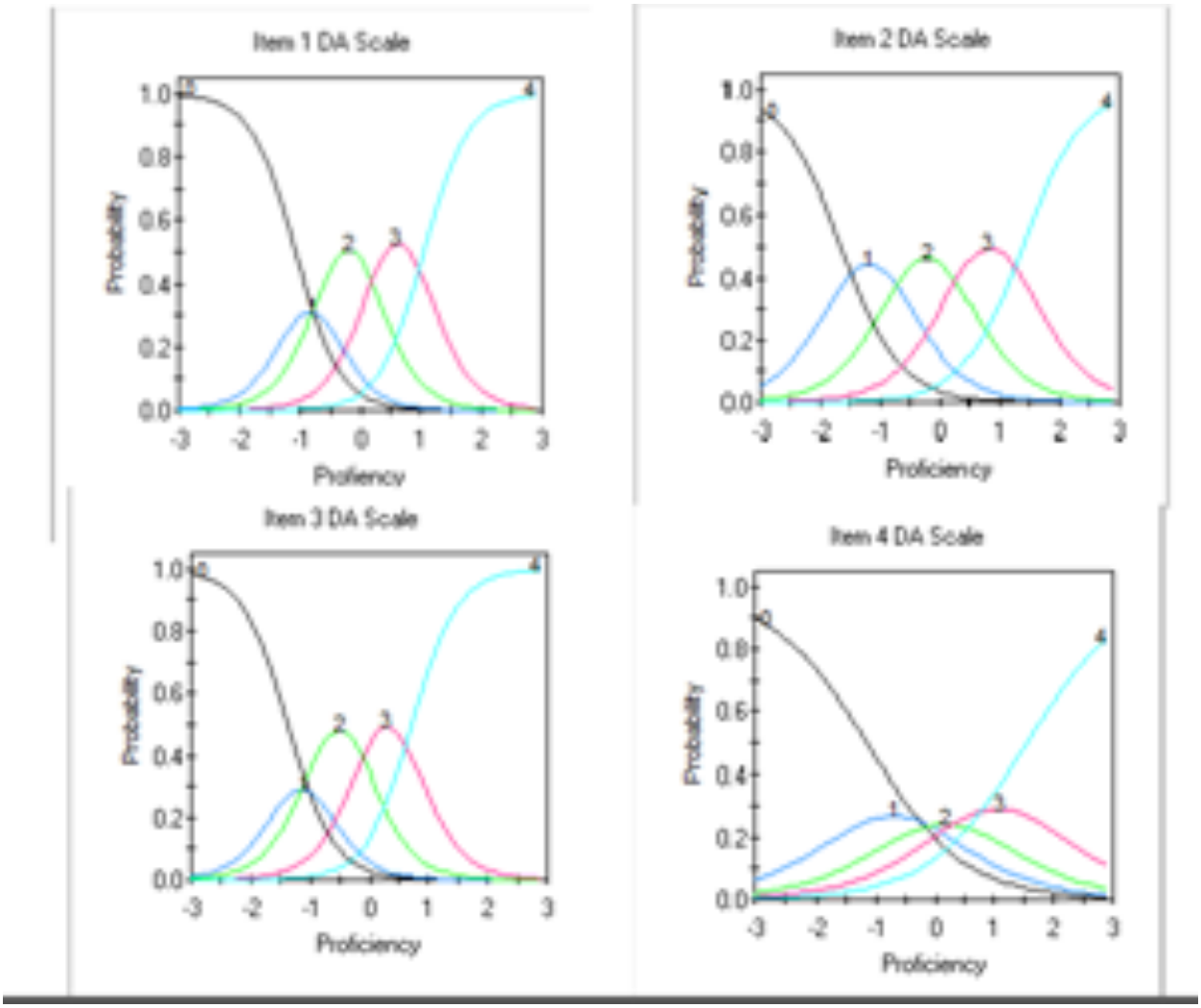

Figure 2. Item Characteristic Curves for TAGG-P DA scale.

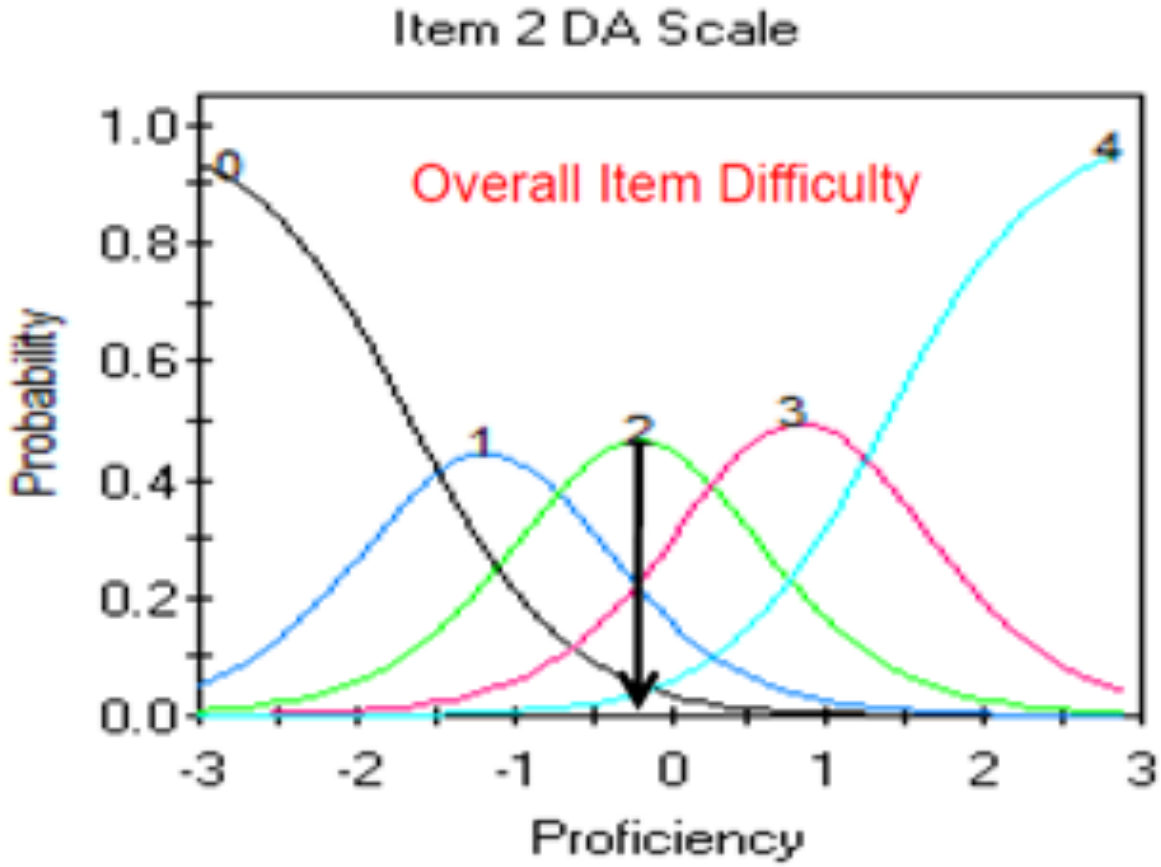

Figure 3. Overall item difficulty for Item 2 TAGG-P DA scale.

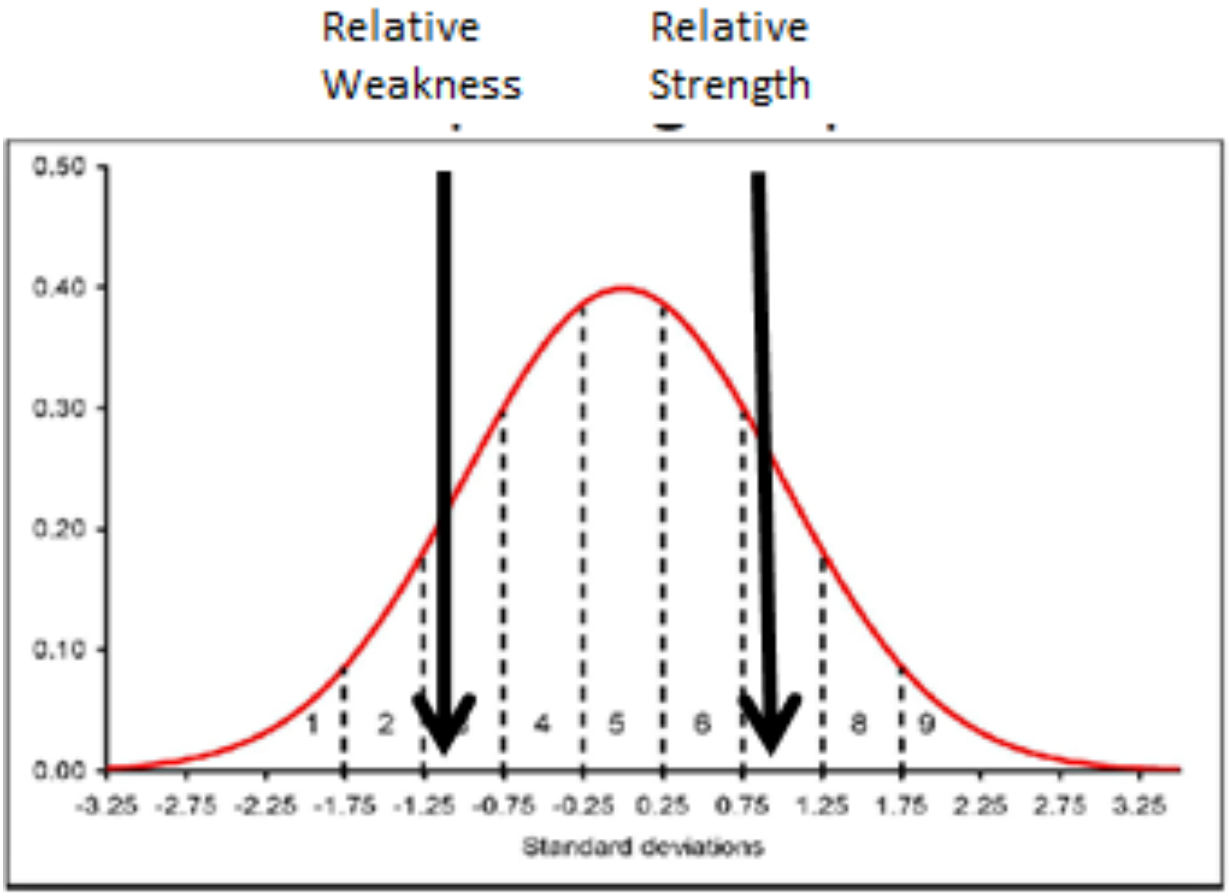

Figure 4. Identifying relative strengths and weaknesses within a student.

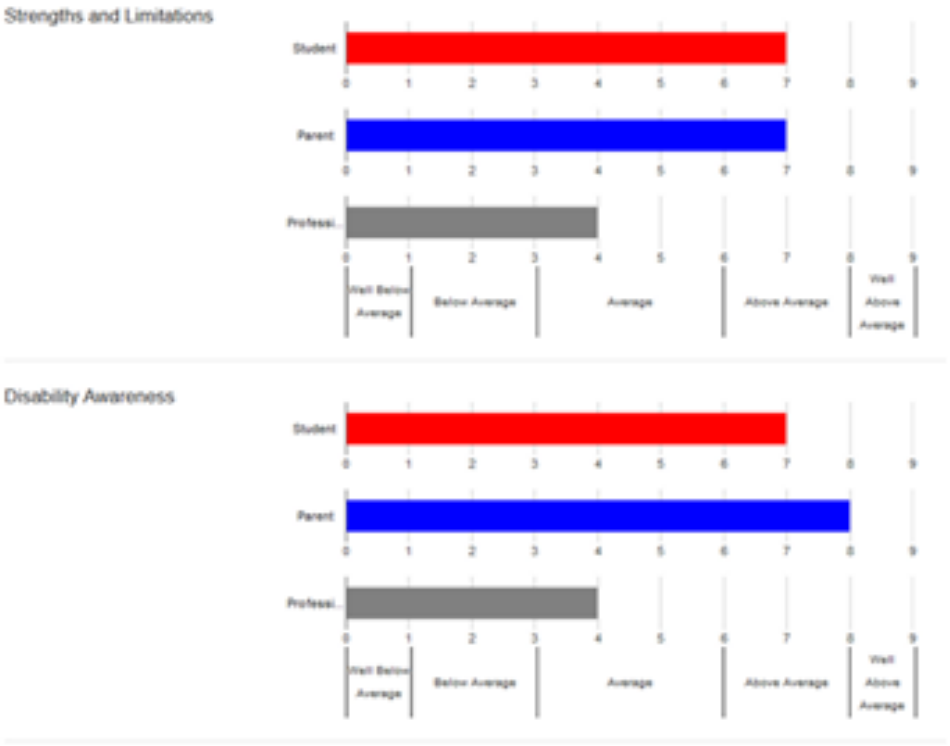

Figure 5. Stanine score presentation.