Z Scores, Standard Scores, and Composite Test Scores Explained

Chittaranjan Andrade[1]

ABSTRACT

Patients may be assessed using a battery of tests where different tests yield scores in different units, where different tests have different minimum and maximum scores, and where higher or lower scores mean different things in different tests. Therefore, a composite test score cannot be obtained by simple addition or averaging of scores in the individual tests. However, if performances in individual tests are converted to Z scores, the Z scores can be added or averaged to yield a composite score that can be interpreted or processed using conventional statistical methods. This article explains in simple ways how Z scores are calculated, what the properties of Z scores are, how Z scores can be interpreted, and how Z scores can be converted into other standard scores.

Keywords: Statistics, Z score, standard score, composite score, T score, stanine score, sten score

In a hypothetical study, I randomize schizophrenia patients to computer-based cognitive remediation (CR) or television viewing (TV) thrice weekly for three months. I administer five cognitive tasks at the study baseline and, again, at the study endpoint. I wish to determine whether CR improves cognitive task scores more than TV does. One way of doing this is to use statistical tests to compare CR and TV groups, task by task; however, there are several problems associated with this approach. For example, performing five separate statistical tests, one for each cognitive task, increases the risk of a Type 1 (false positive) error.[1] Or, patients in one group may perform better in some tasks and worse in other tasks relative to patients in the other group; so, what should the overall conclusion be? Or, patients in one group may perform better than patients in the other group in all tasks without the results reaching statistical significance for any task; again, what should the overall conclusion be?

Need for Composite Scores

One way to get an overall perspective is to create a composite score. This is easily done in some circumstances; for example, one may reasonably add or average language, science, math, geography, and history marks to get a single composite score in school examinations. Simple addition or averaging of marks is possible because all subjects are treated equally, all subjects are marked from 0 to 100, and all subjects have higher marks indicating better performance. Simple addition or averaging is not possible for cognitive tasks because different tasks have different everyday importance, because different tasks have different minimum and maximum scores, because in some tasks lower scores indicate better performance, and in other tasks, higher scores indicate better performance, and because some tasks yield scores measured in units of time, others yield scores measured as the number of correct responses, and so on. As examples, verbal memory may be more important than visual memory because of its application to everyday life; tests of processing speed are measured in units of time and lower scores indicate better performance; and tests of memory are measured in the number of units correctly recalled and higher scores indicate better performance. So, simple addition or averaging of scores is not possible.

Computing Z Scores

One solution is to first convert the original (raw) scores for each cognitive task into a new score that is described in the same unit for all tasks. The new scores can

[1]Dept. of Clinical Psychopharmacology and Neurotoxicology, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka, India.

HOW TO CITE THIS ARTICLE: Andrade C. Z Scores, Standard Scores, and Composite Test Scores Explained. *Indian J Psychol Med.* 2021;43(6):555–557.

Address for correspondence: Chittaranjan Andrade, Dept. of Clinical Psychopharmacology and Neurotoxicology, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka 560029, India. E-mail: andradec@gmail.com

Submitted: 28 Aug. 2021
Accepted: 29 Aug. 2021
Published Online: 11 Oct. 2021

ACCESS THIS ARTICLE ONLINE
Website: journals.sagepub.com/home/szj
DOI: 10.1177/02537176211046525

then be added or averaged to form a composite score. Conversion of the raw scores into Z scores is one such approach. To do this (for example) for the verbal memory test in the cognitive battery, I would need to perform the following actions with the verbal memory raw scores.

1. Calculate the mean (M) and standard deviation (SD) verbal memory score for the CR and TV groups *combined into a single group*; that is, for the *pooled sample*.
2. Calculate the Z score for each patient; the formula is Z = (x – M)/SD, where x is the patient's verbal memory raw score and M and SD are the estimates from the previous step. Positive Z values indicate scores that are greater than the mean of the pooled sample, and negative values indicate scores that are less than the pooled mean.[2]

Z scores are similarly calculated for each patient for each of the remaining four cognitive tasks. This is done separately for the baseline and endpoint data.

Understanding Z Scores

If we look at the formula for the Z score, we will immediately realize that *the Z score tells us how far above or below the mean an individual's score is, expressed in units of SD*. So, if the M(SD) is 18(4) for the verbal memory scores in the pooled sample, a patient with a verbal memory score of 20 has a Z score of (20–18)/4, or 0.5. That is, the patient's verbal memory score is half an SD above the mean of the sample. Another patient whose raw score is 12 would have a Z score of (12–18)/4, or –1.5; that is, one and a half SDs below the sample mean.

Interpreting and Using the Z Scores

The raw scores were in different units in the different cognitive tasks. Z scores are all in the same unit, that is, SD. The Z score distribution has a mean of 0 and an SD of 1. Z scores are useful because they allow data to be interpreted or used in many ways, as the following examples show:

1. The Z score tells us at a glance how the patient has performed relative to the rest of the sample, something that is not evident from the inspection of a raw score.
2. Because we understand the relationship between M and SD in the normal distribution, and because the Z score is an SD unit, we know that Z scores of 2 and above (either positive or negative) are quite far from the mean and that Z scores of 3 and above (either positive or negative) are so far from the mean as to represent outliers. So, an inspection of Z scores can identify outliers in the sample.
3. Using published tables, such as a table of the area under the normal curve, we can read off the probability of obtaining any individual Z value.
4. If a patient has a Z score of, for example, 1 for verbal memory and a Z score of –0.3 for processing speed, because the unit of Z is the same for both tasks, we can conclude that this patient performed better in the verbal memory task than in the processing speed task. This conclusion would not have been possible from an inspection of the raw scores.
5. Z scores can be added to create a composite score. In the context of the study described at the start of this article, the Z scores for the five cognitive tasks can be added for each patient; this creates a composite cognitive (total) score for the patient. There are two noteworthy points.
   a) For tests where lower scores indicate better performance, Z scores should be multiplied by –1 so that when the Z scores are added, the composite score will correctly indicate the direction of change.
   b) Whereas the individual Z scores are in units of SD, the composite score, created by adding the Z scores for the five tests, is no longer in units of SD. However, if the composite score is divided by the number of tests, we get a composite (average) score that is again a unit of SD.

Z Scores and Composite Scores

When Z scores are added or averaged as described above, each cognitive task receives equal weightage. It is possible to create composite scores in which some tasks are given higher weightage than others, based on preset values for weights. For example, it can a priori be decided that, because verbal tasks are more relevant in everyday life than visual tasks, the verbal memory task should receive twice the weight that the visual memory task receives when computing the composite score. Weights can also be determined and assigned through statistical methods.[3,4]

Once the composite score has been calculated for each patient, the M(SD) composite score can be calculated for CR and TV groups separately at the study baseline and at the study endpoint and then processed using usual statistical methods; this can be done whether the composite score is a total or an average of the Z scores of the individual cognitive tasks.

Standard Scores

Some people find it hard to understand Z scores, especially when values are negative (readers are reminded that Z scores have a mean of 0, and that Z values that are negative indicate scores that are below the mean of the sample). This difficulty can be resolved by converting Z scores into other standard scores. The Z score is one example of a standard score; using simple formulae, Z scores can be converted to other standard scores that have only positive values and other specific properties. An example is the T score which has M=50 and SD=10. Stanine and sten scores are based on the same principle. Stanine (standard nine) scores range from 1 to 9, with a mean of 5 and an SD of 2; sten (standard ten) scores range from 1 to 10, with a mean of 5.5 and an SD of 2. Stanine and sten scores are used in some psychological tests. IQ scores are also standardized; they have a mean of 100 and an SD of 15. Readers may note that Z transformation and other methods of standardization do not change the ranking of the original data.

Computing Z Scores: Reprise

The Z score for an individual measurement can be calculated using the mean and standard deviation of a sample, or of a pooled sample, or of the population, depending on the context in which the Z score requires to be derived and used. If the sample comprises a single group, such as a class of students, the

Z scores are based on the M(SD) of that group. If there are two groups, such as in the study described in this article, Z scores should be calculated based on the M(SD) of the pooled sample. However, when Z scores are interpreted for a single individual on a test for which population norms are available, the population mean and standard deviation are used rather than the sample M(SD).

As an aside, for the study described in this article, why are the Z scores computed for the pooled sample; why cannot the Z scores be computed for each group separately, and then M(SD) Z scores compared between groups? The answer ought to be obvious. Z scores may create new values but do not change the ranks (order) of the raw scores within a group; and these new values have an M(SD) of 0(1) because, as stated earlier, this is a property of Z scores. So, if Z scores are computed separately for CR and TV groups, the M(SD) of the z scores for each group will be 0(1), making comparisons between groups illogical. However, if the CR and TV groups are pooled, the order of the raw scores will change; whereas the M(SD) of the Z scores for the pooled group will be 0(1), the Z scores for the CR and TV group patients will depend on the new order, and the M(SD) of the Z scores for the CR and TV groups will no longer each be 0(1). So, the M(SD) Z scores thus created can now be validly compared between the CR and TV groups. Pooling of groups is done in certain nonparametric tests, as well. For example, in the Mann–Whitney and Kruskal–Wallis tests, the groups are pooled, individual values are ranked, and then the ranks are compared across groups.

Parting Notes

Knowledgeable readers may recognize that the standardized mean difference that is a measure of pooled effect size in meta-analysis and the (standardized) beta coefficient in regression analysis are both based on principles similar to those discussed in this article. A discussion on these, however, is out of the scope of the present article.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Chittaranjan Andrade https://orcid.org/0000-0003-1526-567X

References

1. Andrade C. Multiple testing and protection against a Type 1 (false positive) error using the Bonferroni and Hochberg corrections. Indian J Psychol Med 2019; 41(1): 99–100.
2. Norman GR and Streiner DL. *Biostatistics: The bare essentials*. 4th ed. People's Medical Publishing House, 2014.
3. Song MK, Lin FC, Ward SE, et al. Composite variables: When and how. Nurs Res 2013; 62(1): 45–49.
4. Andrade C. Mean difference, standardized mean difference (SMD), and their use in meta-analysis: As simple as it gets. J Clin Psychiatry 2020; 81(5): 20f13681.