When Proportion Consensus Scoring Fails

Kimberly A. Barchard, Spencer Hensley, and Emily D. Anderson
University of Nevada, Las Vegas

Reference: Barchard, K. A., Hensley, S., & Anderson, E. D. (2012, April). *When proportion consensus scoring fails.* Paper presented at the Western Psychological Association Annual Convention, San Francisco, CA.

Contact Information: Kimberly A. Barchard, Department of Psychology, University of Nevada, Las Vegas, 4505 S. Maryland Parkway, P.O. Box 455030, Las Vegas, NV, 89154-5030, USA, barchard@unlv.nevada.edu

Abstract

For most objectively scored test items, there is one and only one correct answer, and experts all agree on what that answer is. However, for some psychological constructs, experts may disagree about the correct answer, or the answer may vary across context or culture. In those situations, another method is needed to identify the correct answer. One increasingly popular method is proportion consensus scoring (PCS), in which a person's score on an item is equal to the proportion of the norm group who gave that same response.

PCS is controversial (Keele & Bell, 2009; Maul, 2011). The purpose of this paper is to determine whether PCS can be used to identify the correct answers on a test. We used items for which there is one and only one correct answer, so that we could determine objectively whether PCS gives the highest score to the correct answer. We hypothesized that PCS would work well for easy or moderate items, but would not work well for difficult items.

A total of 353 undergraduates completed the Las Vegas Vocabulary Test (Barchard, 2004). This test contains 60 multiple-choice items. First, we calculated objective scores using the dichotomous scoring key (1 = right, 0 = wrong). Next, we grouped the items by difficulty: We sorted the items from easiest to hardest, and divided them into three groups of 20. In this sample, the 20 easiest items had mean scores of .60 or higher, and the 20 most difficult items had mean scores of .265 or lower. Third, we constructed the PCS scoring key. If 20% of the sample selected option A, then all participants who selected option A received a score of .20. Finally, we examined the correlations between objective scores and PCS scores. We averaged these correlations across the 20 items at each difficulty level.

As expected, the correlation between PCS and objective scores decreased as item difficulty increased. The average correlations for the easy, moderate, and difficult items were .999, .796, and .235, respectively. These results demonstrate that PCS scoring does a poor job of identifying the correct answer for difficult items.

Introduction

For most objectively scored test items (e.g., a math problem), there is one and only one correct answer, and experts all agree on what that answer is. Creating the scoring key is easy. However, for some psychological constructs (e.g., emotional intelligence), experts may disagree about the correct answer to particular items, or the answer may vary across context or culture. In those situations, another method is needed to identify the correct answer and create the scoring key. One increasingly popular method is to create the scoring key using the responses from the norm group. This is referred to as consensus scoring.

Several types of consensus scoring exist. For tests of emotional intelligence (Mayer, Caruso, & Salovey, 2000; Mayer, Salovey, Caruso, & Sitarenios, 2003; Zeidner, Shani-Zinovich, Matthews, & Roberts, 2005), proportion consensus scoring is often used. In proportion consensus scoring (PCS), a person's score on an item is equal to the proportion of the norm group who gave that same response. For example, if 35% of respondents selected option C, then everyone who selected C would receive a score of .35.

In general, the tests that use consensus scoring have demonstrated adequate reliability and validity (Mayer et al., 2000; Mayer et al., 2003; Zeidner et al., 2005). Within domains of human interaction, consensus scoring is plausible. For example, emotional knowledge evolves within a general social context, and thus group consensus should be able to identify the correct answers (Mayer, Salovey, Caruso, & Sitarenios, 2001). However, empirical investigations of this matter have not always reached the same conclusion. For example, Keele and Bell (2009) examined item responses to the Changes and Blends tasks on the Mayer-Salovey-Caruso Emotional Intelligence Test (Mayer et al., 2003) and found no clear agreement on responses to the items. Moreover, Geher and Renstrom (2004) argued that PCS may be assessing convergence to popular opinion rather than actual ability.

The purpose of this paper is to determine whether PCS can be used to identify the correct answers on a test. We used items for which there is one and only one correct answer, so that we could determine objectively whether PCS gives the highest score to the correct answer. We hypothesized that PCS would work well for easy or moderate items. Most people would select the correct answer, and so people who selected the correct answer would obtain a high score on that item. Moreover, there would be a high correlation between the PCS scores and objective scores. However, we hypothesized that PCS would not work well for difficult items. Most people would not select the correct answer, and so people who selected the correct answer would not get a very high score on that item. Because of this, there would be a low correlation between PCS scores and objective scores.

Method

Participants

A total of 353 undergraduates (208 female, 145 male) participated in this study in return for course credit. They ranged in age from 18 to 50 (M 19.84, SD 3.28). They identified their ethnicities as follows: 58.4% Caucasian, 12.8% Hispanic, 11.1% Asian, 8.8% African American, 5.7% Pacific Islander, and 3.1% Other. Two people did not identify their ethnicity.

Measure

Las Vegas Vocabulary Test (LVVT Barchard, 2004) is a multiple choice test. There are two sections, each containing 30 items in increasing levels of difficulty. Examples of an easy item and a difficult item are given in Figure 1. Each item on the LVVT was designed to have a single correct answer.

Figure 1
Example Items from the Las Vegas Vocabulary Test

36. Surge	27. Demeritorious
a) Encourage	a) Salacious
b) Drip	<u>b) Opprobrious</u>
c) Twill	c) Portentous
<u>d) Swell</u>	d) Palmary
e) Schooner	e) Ostentation

Analyses

To examine the relationship between objective scores and PCS scores, we first had to calculate objective scores. In objective scoring, a response was scored as 1 if it was correct or 0 if it was incorrect.

Next, we grouped the items by difficulty. For each item, we calculated the proportion of respondents who selected the correct answer according to the objective scoring key. Then we sorted the items from easiest to hardest, and divided them into three groups of 20. In this sample, the 20 easiest items had mean scores of .60 or higher, and the 20 most difficult items had mean scores of .265 or lower. Note that these undergraduate students were performing near chance levels on the difficult items.

Next, we constructed a PCS scoring key. If 20% of the sample selected option A, then all participants who selected option A received a score of .20.

Results

As expected, the correlation between PCS and objective scoring decreased as item difficulty increased. Table 1 shows the correlations for the 20 easy items, Table 2 shows the moderate items, and Table 3 shows the difficult items. The average of the correlations for the three types of items were .999, .796, and .235, respectively. These results demonstrate that PCS scoring does a poor job of identifying the correct answer for difficult items.

Conclusions

Proportion consensus scoring works well for easy items. Most people select the correct answer, and so the correct answer is given a high score. For items with a moderate level of difficulty, PCS does not work quite as well, but performance is still reasonable. It does a pretty good job of identifying the best answer. However, for difficult items, PCS performs poorly. Few people select the best answer, and so the people who do select the best answer are given a low score.

The underlying assumption in the use of consensus scoring is that large samples of individuals converge on correct answers (Legree, 1995). This study demonstrates that this rationale is only applicable to easy and moderate items. For difficult items, an alternative rationale is needed. Future research should explore alternative rationales for proportion consensus scoring, and should examine alternative norm-based scoring procedures.

Table 1
Correlations with Veridical Scoring for Easy Items

Item	Veridical Mean	PCS Correlation
5	.99	1.000
3	.96	1.000
1	.96	1.000
34	.95	1.000
31	.95	1.000
37	.93	1.000
35	.93	1.000
8	.93	1.000
39	.92	1.000
32	.92	1.000
2	.89	.999
6	.84	.999
11	.82	1.000
7	.78	.999
40	.76	.997
9	.75	.997
19	.75	.999
18	.69	.993
38	.65	.995
55	.62	.994
Average	.85	.999

Table 2
Correlations with Veridical Scoring for Moderate Items

Item	Veridical Mean	PCS Correlation
16	.58	.983
54	.57	.989
12	.55	.977
4	.54	.711
33	.53	.667
48	.45	.930
25	.44	.943
49	.44	.965
36	.42	.433
14	.42	.953
17	.41	.859
15	.39	.919
42	.38	.470
10	.38	.918
51	.37	.939
43	.34	.901
45	.33	.392
41	.33	.828
23	.28	.470
59	.27	.673
Average	.42	.796

Table 3
Correlations with Veridical Scoring for Difficult Items

Item	Veridical Mean	PCS Correlation
44	.27	.603
27	.26	.702
26	.26	.715
13	.26	.734
52	.25	.114
22	.25	.413
20	.24	.632
50	.24	.436
46	.24	.166
57	.22	.452
47	.20	.259
30	.20	-.019
24	.19	-.006
28	.19	.135
60	.17	.175
56	.17	.097
21	.16	.008
53	.14	-.357
29	.10	-.291
58	.07	-.265
Average	.20	.235

References

Barchard, K. A. (2004). *Las Vegas Vocabulary Test*. [Unpublished Psychological Test] Available from Kimberly A. Barchard, University of Nevada, LasVegas, 4505 Maryland Parkway, Las Vegas, NV, 89154-5030, barchard@unlv.nevada.edu

Geher, G. & Renstrom, K.L. (2004). Measurement issues in emotional intelligence research. In G. Geher (Ed.) *Measuring emotional intelligence: common ground and controversy* (1-17) Hauppauge, NY: Nova Science.

Keele, S. M., & Bell, R. C. (2009). Consensus scoring, correct responses and reliability of the MSCEIT V2. *Personality and Individual Differences, 47*, 740-747. Doi: doi:10.1016/j.paid.2009.06.013

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-base testing procedures. *Intelligence, 21*, 247-266. doi:10.1016/0160-2896(95)90016-0

Maul, A. (2011). The validity of the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) as a measure of emotional intelligence. *Emotion Review*.

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27*, 267-298. doi:10.1016/S0160-2896(99)00016-1

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1*, 232–242. doi:10.1037//1528-3542.1.3.232-242

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*, 97-105 DOI: 10.1037/1528-3542.3.1.97

Zeidner, M., Shani-Zinovich, I., Matthews, G., & Roberts, R. D. (2005). Assessing emotional intelligence in gifted and non-gifted high school students: outcomes depend on the measure, *Intelligence, 33*, 369-391. doi:10.1016/j.intell.2005.03.001