The Balanced Scorecard for Schools

Present Value, Future Promise

Technical Report

Professor Margaret Wu

15 October 2019

Marshall Cavendish Education

TABLE OF CONTENTS

Introduction 12
The Analysis of *BSfB* Trial Data 13
Constructing Measurement scales 13
Reverse coding of negatively worded questions 13
Merging data from all survey forms 14
Data cleaning process 14
IRT scaling results 14
Item statistics presented in the trial analysis 17
School Reports for Trial Schools 18
Construction of the Final Survey Forms 19

2. INTRODUCTION

The "Balanced Scorecard for Schools" (*BSfS*) program is a school-evaluation and improvement program that administers a large bank of survey questions to key stakeholders of a school. The stakeholders include the school governing board, leaders, staff, students and parents/carers. Respondents' ratings on various aspects of school activities are reported by a number of domains. The survey questions are structured according to five domains. They are:

1. School Environment
2. School Leadership
3. Teaching and Learning
4. Well-Being and Equity
5. Use of Technology

Within each domain, there are a number of subdomains. The subdomains are further represented by a number of indicators. For example, within the domain of School Leadership, there is a subdomain called *Leadership Capability and Effectiveness*. One indicator for this subdomain is 'School Leadership Style'. To measure this indicator, there are eight specific questions in the survey forms. A comprehensive survey map can be found in the documentations of *BSfS*.

The survey questions have been written based on international research and have been reviewed by experts in the relevant fields. The survey questions have been further trialled in a number of trial schools in Australia, to obtain quantitative data that provide the evidence that the surveys are working well and are reliable instruments.

This report describes the steps taken in the quantitative data analysis of the trial data. In addition, this report also documents the steps used to construct the final survey forms.

3. THE ANALYSIS OF *BSFS* TRIAL DATA

The *BSfS* surveys were trialled at the end of 2018 in four schools in Australia. Survey items were divided into 30 survey forms, with 3 forms for the school governing board, 3 forms for parents/carers, 6 forms for school leaders, 6 forms for students, 10 forms for teachers, and 2 forms for non-teaching staff. The survey forms had overlapping survey questions so that all data from the different survey forms could be linked and scaled on the same measurement scales.

Constructing Measurement scales

As the *BSfS* surveys aim to provide measures on various aspects of key school activities, it is important to ensure that the measures constructed are reliable and meaningful. For example, if there is a reported score on the quality of teaching practice (called a construct), it is important to make sure that all questions on the quality of teaching construct are indeed tapping into this construct. That is, there needs to be cohesion (or internal consistency) among the questions for this construct so it is possible to add up the scores on individual questions to arrive at an aggregated score for the quality of teaching practice.

To check on the good measurement properties of each scale constructed, Marshall Cavendish Education (MCE) carried out an 'item response theory (IRT)' scaling process, to check for the internal consistency of the survey items for each indicator in the *BSfS* survey map.

Reverse coding of negatively worded questions

A first step before the IRT scaling process is to ensure that negatively worded questions are *reverse-coded* so that the scores from these questions can be added to those that are positively worded. In the *BSfS* surveys, the following questions have been reverse-coded.

Reverse-coded Survey Questions

Question Code
Q1E42
Q2E1261A
Q1E673
Q1E416
Q1E416A
Q1E673A
Q1E563
Q1E1365
Q1E1287
Q2E276

Merging data from all survey forms

As there were many survey forms and overlapping questions across survey forms, the respondents' data from all survey forms were merged together to form one single large dataset, with responses to the same questions arranged in the same columns of the data file. The scaling process was carried out using this merged data file.

Data cleaning process

Prior to analysing the data, a number of checks were carried out. For example, a check that the allocation of survey questions to indicators was consistent, and that missing responses were coded to NA (not available).

IRT scaling results

In all, 36 scales were constructed for 36 indicators. A typical set of results for an indicator is shown below.

```
============================================================
 D1E1S1 1. Effectiveness of school vision formulation
           Number of questions =  7     Reliability =  0.76
============================================================

 ** Q1E64 **        Forms:  GFA2 PGFA1 LFA3 TFA1 NTFA2
   The Principal's personal vision aligns with the official vision for the
school?
   average score = 3.02   discrimination = 0.86   no. of respondents = 168

 ** Q1E88 **        Forms:  GFA2 LFA3 TFA1 NTFA2
   The school's vision describes its expectations of students?
   average score = 2.8   discrimination = 0.82   no. of respondents = 41

 ** Q1E101 **        Forms:  GFA2 PGFA1 LFA3 TFA1 NTFA2
   The school's vision establishes core values for the school.
   average score = 3.01   discrimination = 0.87   no. of respondents = 161

 ** Q1E95 **        Forms:  GFA3 PGFA2 LFA3 TFC1 NTFA2
   The school's vision is easy to understand.
   average score = 2.86   discrimination = 0.89   no. of respondents = 169

 ** Q1E107 **        Forms:  GFA3 PGFA2 LFA3 TFC1 NTFA2
   The process of developing the school's vision increased the levels of t
rust among members of the school community.
   average score = 2.41   discrimination = 0.9   no. of respondents = 164

 ** Q1E25 **        Forms:  LFA1 TFA1
   How much does the school formally involve students in the process of de
veloping its vision statement?
   average score = 1.2   discrimination = 0.97   no. of respondents = 10

 ** Q2E25 **        Forms:  LFA1 TFB1
   How much does the school formally involve students in the process of de
```

veloping its mission statement?
   average score = 1.2   discrimination = 0.96   no. of respondents = 10

                  ****************** D1E1S1 ******************
 Number of questions = 7
 Reliability = 0.76

======================
 Q1E64   average level 3.02   discrimination 0.86

 Seq   Item   Total   Category   Count   Percent   Pbs   MeanAbilit
y
----- ------- ------- ---------- ------ -------- ------ -----------
-
  1   Q1E64   168   a   58   0.35   0.83   2.4
5
  1   Q1E64   168   b   73   0.43   -0.26   -0.9
3
  1   Q1E64   168   c   25   0.15   -0.42   -2.6
2
  1   Q1E64   168   d   6   0.04   -0.29   -3.8
0
  1   Q1E64   168   e   6   0.04   -0.34   -4.3
4

======================
 Q1E88   average level 2.8   discrimination 0.82

 Seq   Item   Total   Category   Count   Percent   Pbs   MeanAbilit
y
----- ------- ------- ---------- ------ -------- ------ -----------
-
  2   Q1E88   41   a   7   0.17   0.61   3.7
1
  2   Q1E88   41   b   22   0.54   0.18   0.8
1
  2   Q1E88   41   c   9   0.22   -0.51   -1.9
6
  2   Q1E88   41   d   3   0.07   -0.41   -3.1
7

======================
 Q1E101   average level 3.01   discrimination 0.87

 Seq   Item   Total   Category   Count   Percent   Pbs   MeanAbili
ty
----- -------- ------- ---------- ------ -------- ------ ----------
--
  3   Q1E101   161   a   42   0.26   0.79   2.
93
  3   Q1E101   161   b   87   0.54   -0.18   -0.

```
58
  3     Q1E101     161        c           23      0.14   -0.48         -2.
97
  3     Q1E101     161        d            9      0.06   -0.38         -3.
86


======================
 Q1E95    average level  2.86    discrimination  0.89

 Seq    Item     Total    Category    Count   Percent     Pbs   MeanAbilit
y
-----  -------  -------  ----------  ------  --------  ------  -----------
-
  4     Q1E95     169        a           47      0.28    0.68          2.5
8
  4     Q1E95     169        b           74      0.44    0.04          0.1
4
  4     Q1E95     169        c           28      0.17   -0.33         -1.6
9
  4     Q1E95     169        d           18      0.11   -0.54         -3.6
7
  4     Q1E95     169        e            2      0.01   -0.33         -7.1
0


======================
 Q1E107    average level  2.41    discrimination  0.9

 Seq     Item     Total    Category    Count   Percent     Pbs   MeanAbili
ty
-----  --------  -------  ----------  ------  --------  ------  ----------
--
  5     Q1E107     164        a           24      0.15    0.60          3.
43
  5     Q1E107     164        b           47      0.29    0.36          1.
29
  5     Q1E107     164        c           72      0.44   -0.34         -1.
00
  5     Q1E107     164        d           15      0.09   -0.42         -3.
27
  5     Q1E107     164        e            6      0.04   -0.44         -5.
52


======================
 Q1E25    average level  1.2    discrimination  0.97

 Seq    Item     Total    Category    Count   Percent     Pbs   MeanAbilit
y
-----  -------  -------  ----------  ------  --------  ------  -----------
-
  6     Q1E25     10         b            1       0.1    0.67          4.9
8
```

```
  6     Q1E25     10         c             3       0.3    0.41          1.0
8
  6     Q1E25     10         d             3       0.3   -0.07         -1.0
2
  6     Q1E25     10         e             3       0.3   -0.78         -4.0
5


======================
 Q2E25    average level  1.2    discrimination  0.96

 Seq    Item     Total    Category    Count   Percent     Pbs    MeanAbilit
y
-----  -------  -------  ----------  ------  --------  ------  -----------
-
  7     Q2E25     10         b             1       0.1    0.52          3.9
7
  7     Q2E25     10         c             2       0.2    0.43          2.2
2
  7     Q2E25     10         d             5       0.5    0.04          0.2
4
  7     Q2E25     10         e             2       0.2   -0.87         -4.0
7
```

Item statistics presented in the trial analysis

A comprehensive list of item statistics was presented in the trial analysis report. These included

- the number of survey questions contributing to a scale for an indicator
- the reliability of the scale
- the average score of respondents (between 0 to 4) on the scale
- the point-biserial correlation index for each survey question
- the number of respondents who answered questions on this scale
- a breakdown of respondents by response categories (a to e)

The trial analysis report was sent to the project personnel of *BSfS*. After reviewing the item statistics, a meeting was held to discuss the trial analysis results. A number of items deemed not 'fitting' with other items on the same scale were removed. Some items had been edited in terms of wording. A number of items had been moved from one scale to another.

As a result of the trial analysis, the *BSfS* survey questions and indicators have been sharpened to better reflect the descriptions for each indicator.

4. SCHOOL REPORTS FOR TRIAL SCHOOLS

In appreciation of the effort and support from the schools who undertook the trialling of the survey questions, a school report was written for each school in the trial. The school reports presented results by domains and subdomains. For each domain, the respondents' scores were aggregated for each of the six stakeholder groups:

- Governing Body (G)
- Leadership Group (L)
- Non-teachers (N)
- Parents/Carers (P)
- Students (S)
- Teachers (T)

The Students' group was further divided into two groups: a "Lower level group" and an "Upper level group", for reporting.

Graphical presentations were made for

- comparisons of domains for each stakeholder group
- comparisons of stakeholder groups within each domain

In addition, the results were also presented in table forms, showing the mean score for each stakeholder group on each domain and subdomain, and the standard deviations of the scores. A graphical plot of the scores also provided a visual impression of the spread of the scores.

To compute an average score for each domain and a stakeholder group, each respondent's responses for that domain were averaged to obtain an average respondent score. The respondent scores for all respondents in a stakeholder group were then averaged to obtain an average domain score for that stakeholder group. In this way, each respondent would contribute the same weight to the average group score. Respondents who took more survey questions would not unduly have more weight when the group average was calculated.

In addition to the trial carried out in Australia, the survey forms were also trialled in Saudi Arabia. The trial responses from Saudi Arabia were not used for the trial analysis, but school reports were generated for the four trial schools in Saudi Arabia.

5. CONSTRUCTION OF THE FINAL SURVEY FORMS

The construction of the final survey forms was based on a number of key criteria:

Firstly, the forms needed to have similar average scores.

For example, to the question

How important is the quality of your workplace to you?

most respondents answered highly important, so the question had a high average score. On the other hand, to the question

How often is classroom air quality discussed at staff meetings?

most respondents answered rarely, so the question had a low average score.

As a consequence, when constructed as a set of questions in a survey form, the expected average score for a form will be higher if it contains a number of questions where the respondents can easily agree, and lower if the questions are more 'difficult' to agree with. In constructing the final survey forms, care was taken to ensure that each survey form had some questions that were easy to agree with, but also questions that were more difficult to agree with. The information about a question's "easiness/difficulty"" could be obtained from the trial analysis. As a result, all final survey forms have similar expected average scores.

Secondly, in constructing the final forms, it is important that each survey form contains questions from all domains. In principle, it will be better to have each respondent answering questions from all domains, than to have respondents answering questions from only one or two domains. In this way the chance of getting biased views from a sub-group of respondents on a particular domain can be reduced

Thirdly, the survey forms have common questions so when analyses are carried out in the future there are linkages across all survey forms so that questions can be scaled on the same scales.