Statistical techniques for studying anomaly in test results: a review of literature

March 2018

Ofqual/18/6355/5

Authors

Qingping He, Michelle Meadows and Beth Black

(Research and Analysis, Office of Qualifications and Examinations Regulation)

Table of Contents

Summary

Anomaly in test results refers to deviation of item response patterns and/or test scores for individual test-takers or groups of test-takers from those that are expected based on theoretical/empirical models or those from others in the sample or the population. There are many factors that can cause anomaly in responses and test scores. These include inappropriateness of the test for the test-takers in terms of the levels of ability of the test-takers and the type of knowledge and skills being assessed by the test; unconventional behaviours of the test-takers in answering questions; inappropriate behaviours such as cheating by either the test-takers themselves or those acting on behalf of them; and others. The existence of aberrance in test results can make test scores inaccurate and invalidate their proposed interpretations and uses. This report reviews a selection of statistical techniques that have been widely used to study anomaly in test results at both individual test-taker and group levels. Particular attention has been paid to the suitability of the various methods to analyse tests of different formats:

- tests composed of multiple choice questions, short-answer questions, or extended-response questions
- linear or adaptive testing

And their features in terms of:

- the underlying assumptions made about the statistical models or empirical relationships used to derive the necessary statistics to measure aberrance
- the power and accuracy of the aberrance measures in detecting anomaly associated with different types of inappropriate behaviours such as answer copying, answer changing, item pre-knowledge, or inappropriate marking/scoring
- and their implications for practical application in terms of interpretability of the aberrance measures, resource requirement, and availability of software packages to conduct the necessary analyses

The report is intended for use as a reference by researchers working in the field of educational assessment.

1. Introduction

Responses to items in a test and the aggregated test score from a test-taker can be affected by a variety of factors, including the appropriateness of the test for the test-taker in terms of the level of ability and the type of knowledge and skills that the test is intended to measure. When a test is appropriate for a test-taker with a certain level of ability, the item response pattern or test score can be predicated with a certain degree of certainty based on the underpinning measurement model used or the response patterns and scores from test-takers with similar level of ability in the sample or population. If an observed response pattern and test score for a test-taker do not conform to the expected pattern and score or those from other test-takers, the response pattern and score are aberrant. There are many factors that can generate anomalous responses and test scores. These include inappropriateness of the test for the test-takers being tested, unconventional behaviours of the test-takers in answering questions, inappropriate behaviours such as cheating by either the test-takers themselves or those acting on behalf of them, and many others (see Meijer, 1996a, b; Karabatson, 2003; Thiessen, 2008). Aberrant responses can result in spuriously high or spuriously low test scores for test-takers. The existence of aberrance in test results can make test scores inaccurate and invalidate their proposed interpretations and uses (see Cizek and Wollack, 2017).

Meijer (1996a, b) described five factors that can cause a test-taker's responses to items in a test to be aberrant, producing spuriously high or low scores: cheating, careless responding, lucky guessing, creative responding, and random responding (also see Karabatson, 2003; Meijer and Sijtsma, 2001; Thiessen, 2008; Emmen, 2011; Tendeiro, 2013). Cheating refers to behaviours where the test-taker illegitimately obtains the correct answers on items which they are unable to answer correctly through pre-knowledge of the items or copying answers from other test-takers or answers provided by their teachers. Careless responding happens when the test-taker answers certain items in the test incorrectly which they are able to answer correctly. Lucky guessing occurs when the test-taker guesses the correct answers to some test items (such as multiple choice items) which they do not know the correct answer. Creative responding happens when a high ability test-taker obtains incorrect answers to certain easy items due to creative and complicated interpretations of the items. Random responding occurs when the test-taker just randomly select the alternatives of some multiple choice items.

Cheating or inappropriate behaviours can take place at both individual test-taker level and group level involving a large number of test-takers from the same class (or school) or different classes (or schools). Cheating that happens at group level can in some cases involve teachers or other relevant people who assist the test-takers to increase their test scores illegitimately. This type of cheating is referred to as educator cheating (Thiessen, 2008). Group level cheating represents test collusion, which may include teacher cheating, test coaching, either by a classroom teacher or

from a review course, systematic answer sharing during the test, use of harvested items, inappropriate marking or scoring test-takers' work, and others (see Wollack and Mayes, 2011; Belov, 2013). As indicated by Belov (2013), test collusion is not limited by the geographic location (eg room, class, school) and can be extended to support various relations between test-takers (eg from the same test-preparation centre, the same group at a social network). Since results from assessments can be used for purposes such as certification of individuals, selection of individuals for further learning/training programmes, and the accountability of teachers and schools, they can be high-stakes for both individuals (students and teachers) and schools. Cheating represents one of the potential negative consequences associated with high-stakes testing (see Cizek, 1999; Madaus et al., 2009). Analysis of anomaly in test results has been used for various purposes, including providing diagnostic information about students' learning and detection of cheating (Meijer and Sijtsma, 2001; Karabatson, 2003; Meijer and Tendeiro, 2014). A large number of statistical techniques have been developed to study aberrant responses and anomalous test scores, with particular focus on detecting anomaly associated with different types of cheating. Although research on methods used to detect test cheating has been primarily focused on individual test-takers, recent years have seen increasing studies on methods used for detecting test collusion (see Wollack and Maynes, 2011; Belov, 2013). Most of such studies were undertaken by researchers in the United States and the Netherlands.

There has been increasing discussion about inappropriate test-taking behaviours by individuals and institutions in high-stakes national assessments and qualifications used in the UK and interests in methods used to identify institutions with unusual performances in tests and examinations (see, for example, Ofqual, 2012; He and Stockford, 2015). There have been numerous studies involving the use of multiple statistical methods to investigate anomaly in test results to detect cheating and test collusion and reviews of specific types of aberrant statistics (see Meijer and Sijtsma, 2001; Karabatson, 2003; Wollack, 2006; Plackner and Primoli, 2014; Meijer et al., 2015). This report intends to provide a more comprehensive review of the most widely used statistical techniques for detecting anomalous test results associated with different types of cheating at both individual test-taker and group levels, drawing on findings from the most recent research in this area and with a particular focus on the potential of using these techniques in the context of high-stakes national assessments used in the UK.

2. Methodology

Research papers and reports were collected from a range of sources for review, including:

- those published in academic journals and books

- those published on the internet by individual researchers, assessment organisations and other research institutions
- unpublished reports from UK exam boards, assessment organisations and other research institutions

Particular attention of the review has been paid to the suitability of the methods to analyse tests of different formats:

- tests composed of multiple choice questions, short-answer questions, or extended-response questions
- linear or adaptive testing

And their features in terms of:

- the underlying assumptions of the statistical models that are made to derive the necessary statistics used for measuring aberrance
- the power (rates of detection) and accuracy (Type I error rates or false positive rates) of the techniques in detecting anomaly associated with different types of inappropriate behaviours such as answer copying, answer changing, item pre-knowledge, or inappropriate marking/scoring for individual test-takers and groups of test-takers
- the implications for practical application in terms of interpretability of the aberrance measures derived, resource requirement, and availability of software packages to conduct the analyses

Effort was also made to briefly describe most of the important steps involved in deriving and applying the techniques reviewed.

3. Analysis of aberrant item response patterns and test scores for individuals

3.1 Analysis based on person-fit (misfit) statistics

Item response or score patterns from individual test-takers may provide additional information to total scores on a test (Meijer and Tendeiro, 2014). One of the approaches used to study item response patterns is person-fit analysis which generally involves the comparison of the observed response pattern from a test-taker with his/her expected response pattern (Karabatsos, 2003; Meijer and Tendeiro, 2014), and a person-fit statistic is derived to characterise the similarity between the observed and expected patterns. If the observed response pattern conforms to the expected response pattern sufficiently well, the person's response pattern is regarded as reasonable or non-aberrant. If, on the other hand, the person's response pattern departs from the expected pattern substantially, his/her response pattern is said to be aberrant or the person is misfitting. There are generally two approaches that can be used to determine the expected response pattern: the

expected (predicted) response pattern is produced using a theoretical or mathematical model (such as an item response theory model) that characterises the interaction between the person and the items in the test, or the expected response pattern is based on the observed response patterns from all test-takers included in the sample. Person-fit indices associated with the first approach are also termed parametric indices, while those associated with the second approach non-parametric indices.

The basic idea used to derive person-fit statistics is that the item response pattern from a test-taker should reflect the difficulty distribution of the items (see Bishop and Stephens, 2013). A test-taker should have a larger chance to answer an easy item correctly than a harder item. Bishop and Stephens (2013) grouped the methods used to derive person-fit statistics into three categories:

- Likelihood: The likelihood approach examines the likelihood that the test-taker's response pattern agrees with the model predicted item response pattern, with higher maximum value of the likelihood function indicating better agreement.
- Covariance: The covariance approach looks at the degree the test-taker's response pattern diverges from the Gutmman Perfect Pattern. If a test-taker answered all easy items correctly but more difficulty items incorrectly, then his/her score pattern is a "Guttman Perfect Pattern".
- Deviation: The deviation approach examines the sum of the differences (or squares of the differences) between the observed responses and the predicted or expected responses for individual items, with higher values indicating larger deviation of the observed item response pattern from the predicted response pattern.

Meijer and Sijtsma (2001) provided a comprehensive review of a wide range of person-fit statistics. Karabatsos (2003), Thiessen (2008), Meijer and Tendeiro (2014) and Meijer et al. (2015) discussed and applied different person-fit indices in their studies. Some of the widely used indices are discussed below in more detail.

3.1.1 Tests composed of dichotomous items

This section discusses indices used for tests composed of dichotomous items.

Non-parametric person-fit indices

Guttman's G indices

The simplest non-parametric person-fit index is the *G* statistic proposed by Guttman (1950, see also Meijer, 1994; Thiessen, 2008). The dichotomous items in a test are sorted according to their difficulty (eg proportion correct). If the items are paired, the G statistic for person n is defined as the counts of the response pairs that deviate

from the Guttman Perfect Pattern (or the number of Guttman errors) (see Meijer, 1994):

$$G_n = \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} X_{ni}(1 - X_{nj}) \quad (1)$$

where X_{ni} and X_{nj} are scores on items *i* and *j* respectively, and J is the number of items in the test.

A perfect Guttman response pattern will produce a G value of zero. Higher G values would represent a greater departure from the Guttman perfect pattern. Since the value of G is likely to increase with an increase in test length, a normalised statistic that takes into account the maximum number of possible Guttman errors was proposed:

$$G_n^* = \frac{G_n}{r_n(J - r_n)} \quad (2)$$

where r_n is the total score of person n on the J test items. Values of G* are in the range [0,1], with 0 representing perfect Guttman distribution and 1 the reversed Guttman distribution.

The $U3$ ***index***

The non-parametric $U3$ index is also a measure of deviance of score patterns and is defined for person n with an observed score r_n as (see van der Flier, 1982; Meijer, 1994; Karabatsos, 2003):

$$U3 = \frac{\ln(X_n^*) - \ln(X_n)}{\ln(X_n^*) - \ln(X_n')} \quad (3)$$

where:

$$\ln(X_n) = \sum_{j=1}^{J} [X_{nj} \log(\frac{p_j}{1 - p_j})]$$

$$\ln(X_n^*) = \sum_{j=1}^{r_n} [\log(\frac{p_j}{1 - p_j})]$$

$$\ln(X_n') = \sum_{j=r_n+1}^{J} [\log(\frac{p_j}{1 - p_j})]$$

$p_j =$ proportion correct by the test-takers on item *j*.

Values of $U3$ can vary from 0 to 1. Again, when the response pattern is the perfect Guttman pattern, the value is 0. If the response pattern is a completely reversed Guttman pattern, $U3$ will be 1. $U3$ can be standardised to have an asymptotically normal distribution (see Meijer and Sijtsma, 2001; Karabatsos, 2003):

$$ZU3 = \frac{U3 - E(U3)}{\sqrt{Var(U3)}} \tag{4}$$

where $E(U3)$ and $Var(U3)$ are the expectation and variance of $U3$ respectively.

The Caution Indices

The Caution Index C examines the ratio of the covariance between a person's item scores and the item proportion scores and the covariance between the person's item scores on the easiest items and the item proportion scores and is calculated from the following equation (see Sato, 1975; Thiessen, 2008; Karabatsos, 2003):

$$C = 1 - \frac{Cov(X_n, p)}{Cov(X_n^*, p)} \tag{5}$$

where:

$p = (p_1, p_2, ..., p_J)$, item proportion correct vector

$X_n = (X_{n1}, X_{n2}, ..., X_{nJ})$, examinee n's item response (score) vector

X_n^* = examinee n's response vector containing correct responses only for the easiest r_n items.

As can be seen, if the person's response pattern is a perfect Guttman pattern, the value of C will be zero. C is therefore a measure of the degree to which the person's item score pattern departs from the Guttman Perfect Pattern. However, C does not have a fixed upper bound and is difficult to interpret (see Meijer and Sijtsma, 2001). Sato suggested that response patterns with C over 0.50 may be regarded as aberrant (also see Huang 2012). Karabatsos (2003) suggested to use 0.53.

Harnisch and Linn (1981) proposed the Modified Caution Index (MCI) which also is a measure of the departure of the response pattern from the Guttman perfect response pattern:

$$MCI = \frac{Cov(X_n^*, p) - Cov(X_n, p)}{Cov(X_n^*, p) - Cov(X_n', p)} \tag{6}$$

where X'_n = examinee *n*'s response vector containing correct responses only for the most difficult $J - r_n$ items. Values of MCI vary from 0 (perfect Guttman pattern) to 1 (reverse Guttman pattern, see Meijer and Sijtsma, 2001; Meijer and Tendeiro, 2014). A critical value of 0.30 was proposed to identify aberrant response patterns.

Sijtsma's H^T Index

The non-parametric index H_n^T proposed by Sijtsma (1986, also see Karabatsos, 2003; Meijor and Tendeiro, 2014) looks at the covariance between the item response pattern of a test-taker *n* and the covariance of the other test-takers and is defined as:

$$H_n^T = \frac{\sum_{m \neq n} (\beta_{nm} - \beta_n \beta_m)}{\sum_{m \neq n} \min[\beta_n (1 - \beta_m), (1 - \beta_n)\beta_m]} \tag{7}$$

where:

β_n = proportion correct for test-taker n over the J test items

β_m = proportion correct for test-taker m over the J test items,

β_{nm} = covariance of item scores between n and m

Values of H^T range from -1 to 1. Persons with low values of H^T are assumed to have aberrant response patterns. When the covariance between a person's response pattern and those of other test-takers is zero, H^T will be zero. When the covariance is negative, H^T will be negative.

Non-parametric cumulative sum (CUSUM) statistics

When studying the responses of persons taking computer adaptive tests (CATs), Bradlow et al. (1998), van Krimpen-Stoop and Meijer (2000, 2001) and Meijer (2002) proposed the use of the item response theory (IRT) based cumulative procedure to detect mis-fitting persons (see later discussion). This procedure can also be applied to other forms of tests such as computer based linear tests (CBTs) and paper and pencil (P&P) tests. The idea of using *CUSUM* statistics in person-fit studies rests on the fact that aberrant behaviour frequently occurs during just one or more of its segments rather than being manifested during the entire test (Armstrong and Shi, 2009). Armstrong and Shi (2009) presented a cumulative sum approach which is based on the likelihood of two probabilities and does not rely on the use of item response function under the IRT framework. The *CUSUM* statistics are conditioned on the number of correct (NoC) scores. The probability, $p_i(s)$, of a correct response to the *i*th item, given the total number of correct answers *s* on the test, is assumed to be the same for all test takers. This probability can be estimated as the unconditional

empirical probability of a correct response. The probability of a person with aberrant response on the item, $p_i^*(s)$, is represented as a upward or downward shift of $p_i(s)$. The difference between $p_i^*(s)$ and $p_i(s)$ can be tested for significance using the likelihood ratio test for the upwards shift γ_i^U or downwards shift γ_i^L:

$$\gamma_i^U = \ln \frac{g_i^U[p_i(s)]^{x_i}\{1-g_i^U[p_i(s)]\}^{1-x_i}}{[p_i(s)]^{x_i}\{1-[p_i(s)]\}^{1-x_i}}$$

$$\gamma_i^U = \ln \frac{[p_i(s)]^{x_i}\{1-[p_i(s)]\}^{1-x_i}}{g_i^L[p_i(s)]^{x_i}\{1-g_i^L[p_i(s)]\}^{1-x_i}}$$

where:

x_i = the score on item *i*

$g_i^U[p_i(s)]$ and $g_i^L[p_i(s)]$ = the upward shift and downward shift functions meeting the following conditions for a test composed of J items:

$$g_i^U(k) = 1$$
$$g_i^L(0) = 0$$
$$g_i^U[p_i(s)] > p_i(s), \quad 0 < s < J$$
$$g_i^L[p_i(s)] < p_i(s), \quad 0 < s < J$$

$g_i^U[p_i(s)]$ and $g_i^L[p_i(s)]$ can be approximated using quadratic functions:

$$g_i^U[p_i(s)] = r_i^U[p_i(s)]^2 + s_i^U p_i(s) + t_i^U, \quad i = 1,2,..., J$$
$$g_i^L[p_i(s)] = r_i^L[p_i(s)]^2 + s_i^L p_i(s) + t_i^L, \quad i = 1,2,..., J$$

In the above equations, the parameters can be estimated using three points meeting the conditions described above. An aberrant pattern can be identified after multiple responses using two of the *CUSUM* statistics, designated as C_i^U and C_i^L respectively, after answering *i* items:

$$C_i^U = \max(0, C_{i-1}^U + \gamma_i^U), \quad i = 1,2,..., J$$
$$C_i^L = \min(0, C_{i-1}^L + \gamma_i^U), \quad i = 1,2,..., J \qquad (8)$$
$$C_0^U = C_0^L = 0$$

For a given level of significance α, the upper bound UB, h_U, and the lower bound LB, h_L, can be estimated empirically. Respondents with values in any of the element of the *CUSUM* statistics beyond the critical values are classified as aberrant respondents.

A more general situation is that the aberrant behaviour can be associated with an upward ability shift on some items and downward shift for some other items. In this case, the following statistic was proposed:

$$C_i^{LR} = \max(C_i^U) - \min(C_i^L) \qquad (9)$$

The *CUSUM* method based on C_i^{LR} statistic is denoted as $CUSUM_{LR}$. Critical values for the three *CUSUM* statistics can be estimated empirically using Monte Carlo simulations.

Parametric person-fit indices

While the non-parametric person-fit statistics are generally derived empirically based on the observed response pattern of the individual concerned and the response patterns of the other test-takers, parametric indices are based on theoretical item response models. Frequently used IRT models for dichotomous items include the one-parameter logistic (1PL) model (which is the same as the Rasch model mathematically), the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) model (see Hambleton et al, 1991). In IRT, the underlying ability or latent trait of an examinee to be measured by the test and the characteristics of the items in the test are specified, and a mathematical function (item response function – IRF) is used to describe the probability that a person will have a specific score on a particular item given his/her ability and the characteristics of the item. The 3PL model can be expressed as:

$$P_{ijx} = \begin{cases} c_j + (1-c_j)\dfrac{\exp(Da_j(\theta_n - \delta_j))}{1+\exp(Da(\theta_n - \delta_j))} & \text{when } x = 1 \\ (1-c_j)\dfrac{1}{1+\exp(Da_j(\theta_n - \delta_j))} & \text{when } x = 0 \end{cases} \qquad (10)$$

where:

D = 1.7;
a_j = the discrimination parameter of item *j*
c_j = the guessing parameter of item *j*
θ_n = the ability of person *n*
δ_j = the difficulty of item *j*
x = 1 or 0, the score of person *i* on item *j*
P_{ijx} = the probability of person *i* scoring x on item *j*

When the guessing parameter is zero, the 3PL model becomes the 2PL model. When the discrimination parameter is 1, the 2PL model reduces to the 1PL model or the Rasch model (see Rasch, 1960; Wright and Stone, 1979):

$$P_{ijx} = \begin{cases} \dfrac{\exp(\theta_n - \delta_j)}{1+\exp(\theta_n - \delta_j)} & \text{when } x = 1 \\ \dfrac{1}{1+\exp(\theta_n - \delta_j)} & \text{when } x = 0 \end{cases} \qquad (11)$$

The Wright's weighted and unweighted person-fit statistics

Wright and Stone (1979) proposed the unweighted and weighted person-fit statistics (also termed outfit and infit statistics) which are residual based for the Rasch model for dichotomous items:

$$U = \frac{1}{J}\sum_{j=1}^{J}\frac{(X_{nj} - P_{nj1})^2}{P_{nj1}P_{nj0}}$$

$$W = \frac{\sum_{j=1}^{J}(X_{nj} - P_{nj1})^2}{\sum_{j=1}^{J}P_{nj1}P_{nj0}} \qquad (12)$$

Values of U and W can vary from 0 to infinity. When a person's response pattern conforms to that predicted by the Rasch model, these indices will be close to 1. The unweighted U is more sensitive to unexpected responses to items with difficulties that are far from the ability of the person, while the weighted W is more sensitive to unexpected responses to items with difficulties that are close to the ability of the person. For both indices, when the value is less than 1, there is less variability in the response pattern than the model predicted (over-fit). When the value is above 1.0, there is more variability in the observed responses than the model predicted. Both indices follow a chi-square distribution with a mean of 1.0 (see Wright and Panchapakesan, 1969; Wu and Adam, 2007). Views on the values of these fit statistics that can be used to identify misfitting persons vary. Persons with infit statistics in the range from 0.70 to 1.30 are normally regarded as fitting the Rasch model well (Keeves and Alagumalai, 1999; Linacre, 2002). However, some researchers set the range of acceptable values for infit and outfit MNSQs even larger, from 0.60 to 1.40 (Tan and Yates 2007; Wong, McGrath and King 2011). Linacre (2002) suggested that when model fit statistics are above 2.0, the measurement system would be distorted.

The Trabin and Weiss' $D(\theta)$ index

The index $D(\theta)$ proposed by Trabin and Weiss (1983) is similar to the unweighted person-fit statistic proposed by Wright and Stone (1979):

$$D(\theta) = \sum_{s=1}^{S} \frac{1}{J_s} \left(\sum_{j \in s} [X_{nj} - P_{nj1}]^2 \right) \qquad (13)$$

where:

J_s = the number of items in subset s

S = number of item sets in the test

Instead of using individual items to calculate the fit statistics, the items are grouped based on their difficulty when calculating the fit statistics.

The likelihood indices

The indices W, U, and $D(\theta)$ discussed above are based on the difference between the observed and the IRT model predicted response patterns. There are other approaches that have been proposed to investigate person model fit. In IRT or Rasch modelling, given the response patterns from a group of persons and the IRT model, both person and item parameters need to be estimated, and this normally involves the use of the likelihood function. In the case of the logistic models for dichotomous items, if the unidimensionaility and local independence assumptions of the models are met (see Hambleton et al., 1991), for a given response pattern $X_n = (X_{n1}, X_{n2}, ..., X_{nJ})$, the likelihood of a person with ability θ to get this response pattern will be $L_n(\theta) = \prod_{j=1}^{J} P_{nj1}^{X_{nj}} P_{nj0}^{1-X_{nj}}$. The log value of the likelihood, $l_n(\theta)$, is:

$$l_n(\theta) = \ln(L_n(\theta) = \sum_{j=1}^{J} [X_{nj} \ln P_{nj1} + (1 - X_{nj}) \ln P_{nj0}]$$

Once an observed response pattern is given, the ability θ of the person can be estimated by maximising the log likelihood $l_n(\theta)$ if the item parameters are known. For the same ability or total test score, different response patterns will produce different maximum values of $l_n(\theta)$ (l_0):

$$l_0 = \ln[L_n(\theta_n)] = \sum_{j=1}^{J} [X_{nj} \ln P_{nj1}(\theta_n) + (1 - X_{nj}) \ln P_{nj0}(\theta_n)] \qquad (14)$$

A higher maximum value of the likelihood will indicate that the response pattern conforms better to that predicated by the IRT model than a response pattern with a

lower value. Levine and Rubin (1979) suggested to use l_0 as a person-fit index. When the IRT model is the 1PL model or the Rasch model, Molenaar and Hoijtink (1990, also see Meijer and Sijtsma, 2001) showed that:

$$l_0 = \sum_{j=1}^{J} \ln[1+\exp(\theta_n - \delta_j)] + r\theta_n - \sum_{j=1}^{j} X_{nj}\delta_j$$

This is because for the Rasch model, the total score is a sufficient statistic. Only the last term in the above equation is influenced by the person's response pattern. Therefore, it was proposed as a fit statistic:

$$M = -\sum_{j=1}^{J} X_{nj}\delta_j \qquad (15)$$

M is easy to calculate. Molenaar and Hoijtink (1990, also see Meijer and Sijtsma, 2001) proposed three approximations to the distribution of M: (1) complete enumeration; (2) Monte Carlo simulation; and (3) a χ^2 distribution, in which the mean, standard deviation (SD), and skewness of M are taken into account.

Drasgow et al (1985) suggested a scandalised form of l_0, l_z, which is approximately asymptotically standard normal:

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} = \frac{l_0 - \sum_{j=1}^{J}[P_{nj1}\ln P_{nj1} + (1-P_{nj1})\ln(1-P_{nj1})]}{\sqrt{\sum_{j=1}^{J} P_{nj1}(1-P_{nj1})[\ln(P_{nj1}/(1-P_{nj1})]^2}} \qquad (16)$$

Since it is a standard normal distribution, when $|l_z \geq 2|$, the response pattern can be regarded as aberrant. Values less than -2.0 indicate that there is more variability in the observed response pattern than the model predicted (under-fit). In contrast, values larger than 2.0 suggest there is less variability in the person's response pattern the model predicted (over-fit). Meijer and Sijtsma (2001) suggested to use a value of less than -1.65 for l_z to flag aberrant response patterns.

Molenaar and hoijtink (1990) and Reise (1995) proved that when the true ability values are replaced with the ability estimates based on the sample, the asymptotic distribution of l_z is not a standardised normal distribution (also see Armstrong et al., 2007). Snijders (2001) modified l_z to make an index denoted as l_z^* which is asymptotically standard normal distribution with sample ability estimates. Magis et al. (2012) suggested that the main reason that l_z^* has not been widely used is because

the original paper by Snijders (2001) was technically complicated. They then derived the index in a more accessible manner:

$$l_z^* = \frac{l_0(\hat{\theta}) - E(l_0(\hat{\theta})) + c_J(\hat{\theta}) r_0(\hat{\theta})}{\sqrt{\tilde{V}(l_0(\hat{\theta}))}} \qquad (17)$$

where:

$$\tilde{V}(l_0(\theta)) = \sum_{j=1}^{J} P_j(\theta) Q_j(\theta) \tilde{w}_j^2(\theta)$$

$$c_J(\theta) = \sum_{j=1}^{J} P_j'(\theta) w_j(\theta) / \sum_{j=1}^{J} P_j'(\theta) r_j(\theta)$$

$$\tilde{w}_j(\theta) = w_j(\theta) - c_J(\theta) r_j(\theta)$$

$$r_0(\theta) + \sum_{j=1}^{J} [X_j - P(\theta)] r_j(\theta) = 0$$

$P_j'(\theta)$ = the first derivative of $P_j(\theta)$

$w_j(\theta)$ = the weight to be specified

$r_j(\theta), r_0(\theta)$ = functions dependent on the IRT model and the estimation method used

The corrected index l_z^* takes into account the sampling variability of the ability estimates when calculating the expectation and variance of the likelihood. Sinharay (2015a) suggested ways to improve the performance of l_z^* further.

The extended modified caution indices

Tatsuoka (1984) extended the non-parametric Caution Index to the three parameter logistic model and proposed four indices.

$$ECI1 = 1 - \frac{Cov(X_n, p)}{Cov(P_n, p)}$$

$$ECI2 = 1 - \frac{Cov(X_n, \overline{G})}{Cov(P_n, \overline{G})}$$

$$ECI4 = 1 - \frac{Cov(X_n, P_n)}{Cov(P_n, \overline{G})} \qquad (18)$$

$$ECI6 = 1 - \frac{Cov(X_n, P_n)}{Cov(P_n, P_n)}$$

where:

$p = (p_1, p_2, ..., p_J)$, item vector of proportion correct

$P_n = (P_{n11}, P_{n21}, ..., P_{nJ1})$, vector of probabilities of correct answers by person *n*

$\overline{G}$ = item vector of average of probabilities of correct answers by all persons

As can be seen, the indices were derived by replacing the easiest response vector in the Caution Index with the expected probabilities of correct answers on all the items in the test, and the item proportion correct vector is replaced with the vector of the average modelled probabilities of correct answers or the expected probabilities of correct answers on all items. The expected value of $ECI1$, $ECI2$ and $ECI4$ is zero, while that of $ECI6$ is a function of the person ability. Tatsuoka further standardised these indices by subtracting their expected values and then dividing by their standard errors which can be estimated based on the IRT model used:

$$SECIk = \frac{ECIk - E(ECIk \mid \theta)}{SE(ECIk \mid \theta)} \qquad (19)$$

where:

$E(ECIk \mid \theta)$ = the expected value for the person with ability θ.

$SE(ECIk \mid \theta)$ = the standard error

k = 1, 2, 4, or 6.

The use of the standardised indices takes into consideration the fact that the original indices tends to inflate values at the extreme values of the ability parameter θ as the error at the extremes are larger than those in the middle of the ability range. Further, they can be interpreted more easily when investigating the nature of aberrant responses.

Parametric cumulative sum (CUSUM) statistics

Bradlow et al. (1998), van Krimpen-Stoop and Meijer (2000, 2001, see also Egberink et al., 2010) and Meijer (2002) proposed the use of the cumulative sum procedure to detect mis-fitting persons in computer adaptive testing (CAT). Assume that $p_i(\theta)$ is the probability of a person with ability θ obtaining a correct answer on the *i*th dichotomous item in the test (for example, the response function representing the 3PL/2PL/Rasch model), a simple statistic T defined as a weighted value of the residual can be used as a measure of inconsistency (see Meijer, 2002):

$$T_i = \frac{1}{J}[x_i - p_i(\theta)]$$

where J is the total number of items in the test and x_i (0 or 1) is the observed score. Two cumulative statistics can be defined as the sum of the T statistic:

$$C_i^+ = \max[0, T_i + C_{i-1}^+]$$
$$C_i^- = \min[0, T_i + C_{i-1}^-] \quad (20)$$
$$C_0^+ = C_0^- = 0$$

Where $\{C_i^+\}$ and $\{C_i^-\}$ reflect the sum of the consecutive positive and negative average residuals respectively. If some appropriate upper and lower bonds, UB and LB, can be established, a response pattern can be classified as aberrant or unlikely if any of the element in $\{C_i^+\}$ and $\{C_i^-\}$ is above or below the bounds:

$$C_i^+ \geq UB, \quad i = 1,2,..., J$$
$$C_-^+ \leq LB, \quad i = 1,2,..., J \quad (21)$$

The upper and lower bonds can be established using simulations or empirically based on the obtained response data (see Meijer, 2002). van Krimpen-Stoop and Meijer (2002) extended the *CUSUM* procedure for tests composed of polytomous items (see later discussion).

3.1.2 Tests composed of polytomous items

While person-fit analysis involving test composed of dichotomous items has been carried out extensively, there has been considerably less research involving tests composed of polytomous items, particularly in the area of using person-fit analysis to identify aberrant respondents. Both non-parametric and parametric approaches have been used to study misfitting persons.

Non-parametric indices

The generalised G indices for polytomous items

Emons (2008) discussed how the concept of Guttman error for dichotomous items could be extended for polytomous items. It is assumed that a score on an item is associated with the number of steps that have been successfully past when answering the item. The introduction of the concept of item steps makes it possible to transform a response or score on an item into an item response vector composed of scores on a series of dichotomous items. The number of elements in the item response is the maximum available score on the item M. Each element in the item response vector represents a step. If a score on the item is m, then the vector will have 1s for the first m elements and 0s for the M-m elements. When all the item response vectors are added for a person, an overall response vector is created. To develop the Guttman index for these dichotomous items, the steps in all the items must be ordered according to their difficulty to form the final response vector for the person. The difficulty of a step in an item is defined as the proportion of respondents with a score equal to or higher than the step score on the item. The item steps within

an item are always ordered. Once the item steps are ordered, the Guttman index (the number of Guttman errors) for person *n* (G^P) can be calculated in the same way as that used for dichotomous items:

$$G^P = \sum_{i=1}^{J_{M-1}-1} \sum_{j=i+1}^{J_M} y_i(1-y_j) \qquad (22)$$

where:

y_i = the value of element *i* (or step) in the response vector $y = (y_1, y_2, ..., y_{J_M})$

J_M = the total number of item steps or the length of the response vector y

When the response pattern is a Perfect Guttman Pattern (ie the easiest items were answered correctly, without any partial scores), then the value of G^P is 0. When all the items in the test are dichotomous items, G^P reduces to the index for dichotomous items.

Emons (2008) also extended the normalised Guttman index for dichotomous items to polytmous items by dividing G^P by the maximum possible Guttman errors $\max(G^P \mid X_+)$ for a given test score $X_+ = \sum_{i=1}^{J_M} y_i$:

$$G_N^P = \frac{G^P}{\max(G^P \mid X_+)} \qquad (23)$$

The values of G_N^P vary from 0 (perfect Guttman pattern or no misfit) to 1 repsenting extreme misfit. Emons suggested that since the item steps in the response vector are structurally dependent, $\max(G^P \mid X_+)$ cannot be expressed in closed form. He developed a recursion algorithm to estimate it.

The generalised $U3$ person-fit statistic $U3^P$ for polytomous items

Emons (2008) also proposed a generalised form of the $U3$ index for polytomous items ($U3^P$) which is defined as follows:

$$U3^P = \frac{\max(W \mid X_+) - W(y)}{\max(W \mid X_+) - \min(W \mid X_+)} \qquad (24)$$

where:

$$W(y) = \sum_{k=1}^{J_M} y_k \log(\frac{\hat{\pi}_k}{1-\hat{\pi}_k})$$

$\hat{\pi}_k$ = the difficulty of step *k*

Values of $U3^P$ vary from 0 (suggesting perfect fit) and 1 (indicating extreme misfit). As with the calculation of the normalised Gutmman index for polytomous items, $\max(W \mid X_+)$ and $\min(W \mid X_+)$ cannot be expressed in closed form and can be calculated using a recursion method (Emons, 2008).

Parametric indices

The Rasch model was originally developed to analyse tests composed of dichotomous items (see Rasch, 1960) and has been extended subsequently for analysing polytomous items. These extended Rasch models include Andrich's Rating Scale Model (RSM), Masters' partial credit model (PCM), and other models (see Andrich, 1978; Masters, 1982; Wright and Masters, 1982; Muraki, 1992). The PCM states that, for a polytomous item with a maximum available score of m (the number of score categories minus 1), the probability $P(\theta, x)$ of an examinee with ability θ scoring *x* on the item can be expressed as:

$$P(\theta,x)=\begin{cases}\dfrac{\exp\sum_{k=1}^{x}(\theta-\delta_k)}{1+\sum_{l=1}^{m}\exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x=1,2,...,m\\[2ex] \dfrac{1}{1+\sum_{l=1}^{m}\exp[\sum_{k=1}^{l}(\theta-\delta_k)]} & \text{for } x=0\end{cases} \qquad (25)$$

where δ_k is the location of the k^{th} step on the latent trait continuum and is referred to as the item step parameter associated with a score category (also frequently referred to as step difficulty or threshold). However, δ_k cannot be interpreted as the difficulty of scoring a score of k on the item. $P(\theta, x)$ is also frequently referred to as category response function (CRF) or item category probability function (CPF). The step parameter δ_k represents the location of the score category on the ability continuum beyond which the probability of achieving a score of k is higher than that achieving a score of $k-1$. The PCM reduces to the Rasch model for dichotomous when the number of response categories is two.

The generalized partial credit model (GPCM) proposed by Muraki (1992) represents an extension of the 2PL model for dichotomous items to polytomous items. The model is also an extension of the Maters' partial credit model by introducing a discrimination parameter for items. In the GPCM, the probability $P(\theta, x)$ of a test-taker with ability θ scoring *x* on the item can be expressed as:

$$P(\theta, x) = \frac{\exp \sum_{k=0}^{x} a(\theta - \delta_k)}{\sum_{l=0}^{m} \exp[\sum_{k=0}^{l} a(\theta - \delta_k)]} \qquad x = 0,1,2,..., m \tag{26}$$

where *a* is the item discrimination parameter. When the maximum score on the item *m* is 1, the GPCM becomes the two-parameter logistic model. When the item discrimination parameter is 1, the GPCM reduces to the PCM.

As suggested by Sung and Kang (2006), the graded response model (GRM) proposed by Samejima (1969) can also be viewed as a generalization of the 2PL model for dichotomous. The model uses the 2PL item response function to model boundary characteristic curves across score categories or the cumulative probability of a response higher than a given category *x*. The probability of scoring a specific score *x*, $P_x(\theta)$, on the item is then calculated as the difference between the cumulative probabilities of achieving the score below the specified score $P^*_{x-1}(\theta)$ and the specified score $P^*_x(\theta)$:

$$P_x(\theta) = P^*_{x-1}(\theta) - P^*_x(\theta)$$
$$P^*_x(\theta) = \frac{\exp[a(\theta - \delta_x)]}{1 + \exp[a(\theta - \delta_x)]} \tag{27}$$

As indicated by Sung and Kang (2006), the GRM is different from the GPCM and PCM in that it requires a two-step process to compute the conditional probability for a test-taker responding in a particular category.

The standardised likelihood index l_z^p for polytomous items

The standardised likelihood index for dichotomous IRT models discussed above can be extended to polytomous IRT models such as those presented above. Drasgow et al. (1985) provided a general form of the index for polytomous items (see also Sinharay, 2015b):

$$l_z^p = \frac{l_0^p - E(l_0^p)}{\sqrt{\text{var}(l_0^p)}} \tag{28}$$

where:

$$l_0^p = \sum_{j=1}^{J}\sum_{k=1}^{m_j}\delta_j(k)\log[P_{jk}(\theta)]$$

$$E(l_0^p) = \sum_{j=1}^{J}\sum_{k=1}^{m_j}P_{jk}(\theta)\log[P_{jk}(\theta)]$$

$$\mathrm{var}(l_0^p) = \sum_{j}\left[\sum_{k}\sum_{l}P_{jk}(\theta)P_{jl}(\theta)\log\left(P_{jk}(\theta)\right)\log\left(P_{jk}(\theta)/P_{jl}(\theta)\right)\right]$$

Where m_j is the maximum score of item *j*, and $\delta_j(k)$ is an indicator function which is 1 if *k*=*j* and 0 otherwise. l_z^p will be asymptotically normally distributed. Sinharay (2015b) recently discussed this statistic for tests containing a mixture of dichotomous and polytomous items.

Other person-fit statistics for the Rasch model

For the partial credit model, person-fit statistics similar to those for the Rasch model based on residuals can also be derived. The weighted and unweighted person-fit statistics are defined as follows:

$$U^p = \frac{1}{J}\sum_{j=1}^{j}\frac{(X_{nj}-E_{nj})^2}{W_{nj}}$$

$$W^p = \frac{\sum_{j=1}^{J}(X_{nj}-E_{nj})^2}{\sum_{j=1}^{J}W_{nj}} \qquad (29)$$

where the expected score on the item and the variance are calculated from:

$$E_{nj} = \sum_{k=0}^{m_j}kP_{njk}(\theta)$$

$$W_{nj} = \sum_{k=0}^{m_j}(k-E_{nj})^2P_{njk}(\theta)$$

These person-fit statistics have been widely used as general diagnostic tools to assess person model fit (see Wright and Masters, 1982).

Cumulative sum (*CUSUM*) *statistics*

van Krimpen-Stoop and Meijer (2002) and van Krimpen-Stoop et al. (2010) extended the *CUSUM* procedure for tests composed of polytomous items used in computer adaptive tests. The residual for a person with ability θ on a polytomous item *i* with $m+1$ response categories contained in a test with J items is calculated from:

$$T_i = \frac{1}{J}\left[x_i - \sum_{j=0}^{m} jP(\theta, j) \right] \quad (30)$$

where x_i is the observed score and $P(\theta, j)$ is the probability of scoring a score of *j* on the item with a maximum score of *m* described by a polytomous IRT model such as the PCM and GPCM. The *CUSUM* statistics defined for dichotomous items discussed previously can be similarly defined for polytomous items and used for identifying aberrant response patterns.

A person-fit index derived from factor analytic models

Ferrando (2007, 2009; also see Clark, 2012; Clark et al., 2014) discussed the use of factor analytic models to study misfitting respondents. For a one-factor analytic model, the score on item *j* from person *n*, X_{nj}, is modelled using a linear function:

$$X_{nj} = \mu_j + \lambda_j \theta_n + \varepsilon_{nj}$$

where μ_j and λ_j are the item parameters, and θ_n is a factor score and ε_{nj} is the random error with variance σ_ε^2. The expected score of X_{nj} on the item is $\mu_j + \lambda_j \theta_n$. A residual based person-fit statistic can be defined as:

$$lco_n = \sum_{j=1}^{J} \left[\frac{X_{nj} - \mu_j - \lambda_j \theta_n}{\sigma_{\varepsilon j}} \right]^2 \quad (31)$$

When sample estimates are used, lco is distributed as χ^2 with J-1 degree of freedom. The one-factor analytic model was subsequently extended to multiple factors. For a K-factor model, a person-fit statistic can be similarly defined:

$$Mlco_n = \sum_{j=1}^{J} \left[\frac{X_{nj} - \mu_j - \lambda_{j1}\theta_{n1} - \ldots - \lambda_{jK}\theta_{nK}}{\sigma_{\varepsilon j}} \right]^2 \quad (32)$$

$Mlco_n$ is also χ^2 distributed, with the number of freedom being J-K. Clark et al. (2012, 2014) discussed the potential application of the difference in person-fit statistics between the one-factor model and a two-factor model as a person-fit statistic:

$$Mlco_{n,diff} = lco_n - Mlco_n \quad (33)$$

This statistic should also be χ^2 distributed with a degree of freedom of 1. Clark et al. (2014) used this statistic to detect cheating due to prior knowledge of portions of items in a test. They argued that if a subset of items in a test had become

compromised and a subset of test-takers took the test with prior knowledge of these items, additional covariance amongst these compromised items for cheating test-takers may result in improved fit at the person level for individuals who engaged in misconduct when a second factor is added to the initial unidimensional model.

3.1.3 Performance of person-fit statistics

The power (the detection rate) and accuracy (Type I error rate or false positive rate) of an aberrant detection statistic under a nominal Type I error rate or α level defined by a theoretical or empirical critical value are normally studied using simulations under different conditions (see Meijer and Sijtsma, 2001; Karabatsos, 2003; Thiessen, 2008). The power of a person-fit index can be affected by a number of factors such as the ability distribution of the test-takers involved, the difficulty distribution of the items, the discrimination distribution of the items, test length, the types of aberrant responses, the proportion of misfitting items, and the proportion of aberrant respondents (see Meijer and Sijtsma, 2001; Karabatsos, 2003; Emons, 2008; Tendeiro and Meijer, 2013; Meijer and Tendeiro, 2014). Meijer and Sijtsma (2001) and Meijer and Tendeiro (2014) also indicated that, the power of person-fit statistics to detect aberrant response patterns increases with increasing item discriminations, test length, and a large spread of item difficulties.

Using simulations, Meijer (1994) demonstrated that the power of the G indices varied from 24% to 100% under different simulation conditions. He found that these Guttman person-fit statistics can be as powerful as or even better than some more complex person-fit statistics in detecting aberrance associated with cheating and guessing. Both G and G^* are easy to calculate and interpret. Thiessen (2008) suggested that the disadvantages of the Guttman's G statistics are that each Guttman error is given equal weight and there is no consensus as to the critical values that should be used to classify aberrant score patterns.

Using the 3PL model and simulations, Drasgow et al. (1987) looked at the performance of a range of person-fit indices in detecting aberrant response patterns and found that the standardised likelihood index l_z is one of the most effective indices in detecting aberrant test-takers, with detection rates varying from 35% to 98% at a false alarm rate of 5%, depending on the simulated conditions. Similarly, using the Rasch model and simulations, Li and Olejnik (1997) compared the performances of a range of person-fit statistics including l_z, the standardised extended caution indices $SECI2$ and $SECI4$, and the standard normal form of the Wright's W and U indices. They found that these indices performed equally well regardless of the type of misfit and test length, with average detection rates varying from 37% for a 30-item test to 51% for a 60-item test. In their study, the Type I error rates or false positive rates were less than the nominal level of 5% used. The l_z index was recommended for detecting spuriously high response patterns. Thiessen

(2007) also used simulation studies to investigate the effectiveness of U, W, l_z and *MCI* in detecting cheaters and found that MCI was able to detect 86% of simulated cheaters. For the other three parametric indices, the detection rates were slightly less, varying from 66-80%. The standardised likelihood index l_z was found to produce the lowest false positive rate. Results from simulations by Armstrong et al. (2007) indicated that the detection power of the l_z index was largely hinged on test characteristics, particularly test difficulty. They therefore suggested that it should be used with caution in an operational testing environment.

Karabatsos (2003) compare the performance of 36 non-parametric and parametric person fit statistics in detecting five types of aberrant response patterns for tests composed of dichotomous items: cheating, careless responding, lucky guessing, creative responding, and random responding under different conditions. The study included different percentages of aberrant examinees and test length. It was found that the sensitivity of these indices was affected by the percentages of simulated aberrant respondents. The most effective five person fit-statistics in detecting aberrant-responding persons were found to be the index H_n^T, the C index, the MCI index, the $U3$ index, and the $D(\theta)$ index. Further, H^T out-performed the parametric $D(\theta)$ and the C and MCI indices. The $U3$ index also outperformed many well-known parametric person-fit statistics. Huang (2012) compared a range of parametric and non-parametric person-fit statistics and found that non-parametric indices performed better than IRT-based parametric indices. Tendeiro and Meijer (2014) recently compared different group-based non-parametric statistics for dichotomous items and concluded that, for a given Type I error rate, H_n^T, followed by $U3$, and C, had generally the highest power to detect misfitting response vectors.

Armstrong and Shi (2009) found that the power of $CUSUM_{LR}$ varied from 47% to 100% in detecting aberrant respondents, depending on the simulation conditions and the specified α level, and the Type I error rates were close to the α values. They also found that the proposed *CUSUM* procedure outperformed considerably other selected model-free non-parametric statistics. The distribution of *CUSUM* statistics can also be used to examine where the aberrant behaviour happened in the response process.

Emons (2008) used simulations to compare the performance of the three non-parametric person-fit statistics G^P, G_N^P and $U3^P$, and the parametric statistic l_z^p for tests composed of polytomous items and found that the detection rate of these indices varied from slightly over 10% to nearly 80% at the 0.05 significance level, depending on the type of aberrant responses and the number of misfitting items. The non-parametric statistics performed almost as well as the parametric statistic in many situations. For a real dataset, he found that the correlations between these non-parametric indices ranged from 0.88 to 0.89.

Using simulation studies, Clark et al. (2014) found that $Mlco_{diff}$ performed better than lco in identifying cheating persons. The detection rate of $Mlco_{diff}$ varied from 12% to 89% at the 5% significance level, depending on the simulated conditions. The Type I error rate was generally small than the nominal value of α. They observed that person-fit statistics like lco measure the difference between observed and expected performance on an item. However, the difficulty of the items which will influence the expected scores is estimated from the observed responses from all test-takers. If a larger proportion of cheaters are present, their influence will make the exposed items to become easier than they should, which will result in smaller residuals when the observed performance on the items is compared with the expected performance. This will reduce the power of residual-based person fit statistics like lco. The lco difference method seems to be more robust compared with the lco approach. They further suggested that increasing the proportion of cheaters can improve performance of the lco difference method when exposed items have a wide range of difficulty, since more cheaters will help produce better estimates for the second factor.

3.1.4 Practical applications and computer software

A large number of person-fit indices have been developed and used operationally. They vary in their power to detect different types of aberrant responses. The situation is further complicated by the fact that even for the same type of aberrant behaviour, there may be several indices available which may perform differently. As indicated by Tendeiro and Meijer (2014), the existence of a large number of person-fit statistics is useful but can also cause confusion as to which of them should be used when a decision to select the best statistics is needed. Tendeiro and Meijer (2014), Meijor et al (2015) and Tendeiro et al. (2016) attempted to provide practical guide for selecting person-fit statistics. The criteria for selecting a person-fit statistic would include:

- high detection power
- lower false positive rates
- interpretability of the critical values
- practicality in terms of resources required to produce the statistic

It is worth noting that, all person-fit statistics, particularly non-parametric person-fit statistics, are generally sensitive to violations against the Guttman model (Meijer and Tendeiro, 2014). IRT based parametric indices will also be sensitive to violations of model assumptions. When comparing the performance of the non-parametric indices and parametric indices, Karabatsos (2003) observed that parametric fit statistic uses the dataset twice, once for the estimation of the model parameters to construct the predicted response patterns and once again to measure its fit to the same predicted response patterns. He suggested that parametric person-fit statistics, based on IRT

model parameters, suffers from this dependence between data and parameter estimates. Non-parametric person-fit statistics on the other hand, circumvent such dependence, which may explain why some of the non-parametric person-fit indices performed better than some of the well-known parametric fit statistics. The advantage of non-parametric approaches is that the underlying non-item response theory model is less restrictive with respect to the data than their parametric counterparts (Emons, 2008). Tendeiro and Meijer (2014) suggested that a high percentage of respondents simultaneously flagged by several person-fit indices could be an indication of aberrant response behaviour.

Tendeiro (2015) has developed an R Package which implements a range of non-parametric and parametric person-fit statistics for tests composed of both dichotomous items and polytomous items. These are listed in Table 1 below.

Table 1 Person-fit statistics available in the R Package PerFit (extracted from Meijer et al, 2015)

Type of statistics	Statistics	References	Type of data	
			Dichotomous	**Polytomous**
Non-parametric	$r.pbis$	Donlon and Fischer (1968)	X	
	C	Sato (1975)	X	
	G, G_n	van der Flier (1980); Meijer (1994)	X	
	A, D, E	Kane and Brennan (1980)	X	
	$U3, ZU3$	van der Flier (1982)	X	
	C^*	Harnisch and Linn (1981)	X	
	NCI	Tatsuoka and Tatsuoka (1982, 1983)	X	
	H^T	Sijtsma (1986)	X	
	G^P	Molenaar (1991)		X
	G_N^P	Molenaar (1991), Emons (2008)		X
	$U3^P$	Emons (2008)		X
Parametric	l_z	Drasgow et al. (1985)	X	
	l_z^P	Drasgow et al. (1985)		X
	l_z^*	Snijders (2001)	X	

3.2 Answer copying and similarity analysis

Answer copying involves a test-taker (the copier) copies answers from another test-taker or other test-takers (the source/s) (see Zopluoglu, 2017). The methods discussed in this section are for test composed of multiple choice questions (MCQs). Most of the statistics used for detecting answer copying are based on the comparison of the amount of overlap or similarity in answers between two test-takers with the normal amount that would be expected if the two test-takers were known to have answered the questions independently of each other. The overlap can be focused on identical incorrect answers or both incorrect and correct answers. If the observed amount of overlap is significantly different from the expected normal amount, copying is assumed to have happened. Most copying indices involve the estimation of the probabilities of the copier to select particular answer alternatives of the items in the test that the source selected. Both CTT and IRT models have been used in deriving copying indices (Wollack, 1997, 2004, 2006; Sotaridona and Meijer, 2003), with the critical values established empirically for CTT and theoretically for IRT. When the copier and the source are not specified, the copying index is referred to the similarity index.

Some indices may be more effective in detecting copying than others, depending on the types and amount of copying. Wollack (2006) suggested there are broadly three types of copying:

- random copying where the copier copies the answers to items randomly from the source(s)
- strings-based copying where the copier copies consecutive strings of items from the source
- mixed copying where a combination of random and strings-based copying is used.

Wollack classified copying indices into two broad categories, depending on the way the responses are used:

- indices that incorporate information from only incorrect responses
- indices that use information from all responses

3.2.1 Answer copying and similarity indices

Angoff's B index and H index

Angoff (1974) proposed to use the B index to study the number of identically incorrect answers between the copier and the source, in comparison with those of test-takers with similar values for the product on incorrect answers between two test-takers (see also Wollack, 2006). The calculation of the B index involves the following steps (Wollack, 2006):

- for the alleged copier and the source, work out the number of identical incorrect items, which is denoted as Q_{ij}, and the product of their number of wrong answers W_iW_j which is used as the conditioning variable
- divide the dataset into strata such that the test-takers within a stratum are homogeneous with respect to the conditioning variable W_iW_j
- for the stratum to which the copier and the source belong, work out the mean of the Q values of all pairs of test-takers and their standard deviation, denoted as $\overline{Q}_{W_iW_j}$ and $S_{Q_{W_iW_j}}$, respectively. The B index is defined as:

$$B = \frac{Q_{ij} - \overline{Q}_{W_iW_j}}{S_{Q_{W_iW_j}}} \tag{34}$$

B therefore assesses the departure of the observed number of identical incorrect answers between the copier and the source from the mean of pairs of test takers with similar values of W_iW_j. It is assumed that B follows the standard normal distribution. Large values of B would be an indication of answer copying.

Another index proposed by Angoff (1974) is the H index which is used to study the magnitude of the maximum number of identical incorrect or omitted items in any string of identical responses in comparison with those of test takers with similar number of omitted or incorrectly answered items. The calculation of the H index involves (see Wollack, 2006):

- for the alleged copier c and the source s, work out the maximum number of identical incorrect or omitted items in any string of identical responses, which is denoted as K_{CS}
- the dataset is partitioned into raw score groups. The group which contains the one with the higher raw score of c and s is used as the comparison group
- for the comparison group, work out the mean of the K values of all pairs of test-takers and their standard deviation, denoted as $\overline{K}_+$ and S_+, respectively. The H index for C and S is defined as:

$$H = \frac{K_{CS} - \overline{K}_+}{S_+} \tag{35}$$

It is also assumed that H follows the standard normal distribution and large values would suggest answer copying.

The K indices

Holland (1986) proposed a statistic, the K-index, to assess the degree of unusual agreement between the incorrect answers on a multiple choice question test of two

test-takers, the copier (c) and the source (s). The following steps will need to be taken to calculate the K index (see Sotaridona and Meijer, 2002):

- determine the group of test-takers with the same number-incorrect score as the copier (subgroup c'). Denote the total number of test takers in the group as $n_{c'}$
- for each test-taker in group c', determine the number of items that match the incorrect answers of the source
- for a copier c in group c', denote his/her number of matched incorrect answers with the source as $m_{c'c}$ and the number of test-takers whose number of matched incorrect answers with the source is greater than or equal to $m_{c'c}$ as n'. The K index for the copier is calculated as the ratio of n' to $n_{c'}$:

$$K_c = \frac{n'}{n_{c'}} \tag{36}$$

That is the K index is defined as the proportion of test takers in subgroup c' whose number of matched incorrect answers with the source is greater than or equal to that of the copier.

The logic of using K_c as an indicator of copying is that when K is very small, there is statistical evidence that test taker c copied from the source s. As Sotaridona and Meijer indicated that the reason for K to be calculated conditional on the number of incorrect scores of the suspected copier is that the number of matching incorrect scores generally depends on the ability levels of c and s. the number of matched incorrect answers will be small when either of the copier or the source or both have high abilities (with many correct scores). When both test-takers have many wrong answers, the matched number of incorrect answers will be large. When the sample size is small (for example less than 100), $n_{c'}$ can become small which will affect the accuracy of the value of the K index. For small samples, Holland (1996) suggested to use the binomial distribution to approximate the distribution of matched number of incorrect answers for a subgroup. For subgroup c', calculate the mean of empirical agreement of incorrect answers between the members in the group and the source:

$$\overline{m}_{c'} = \frac{\sum_{i'=1}^{n_{c'}} m_{c'i'}}{n_{c'}}$$

where $m_{c'i'}$ is the number of matched incorrect answers between test-taker i' in subgroup c'. Assuming that the number of wrong answers of the source is w_s, the average percentage agreement between the subgroup c' and the source is:

$$p^*_{c'} = \frac{\overline{m}_{c'}}{w_s}$$

$p^*_{c'}$ is used as the success probability parameter in the binomial distribution. The corresponding K index, which is denoted as K^*, is defined as the probability of the matched incorrect answers greater than or equal to $m_{c'c}$ and is calculated from:

$$K^*_c = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} p^{*g}_{c'} (1 - p^*_{c'})^{w_s - g} \tag{37}$$

The value of $p^*_{c'}$ is affected by the sample size. When the size is small, its reliability declines. Holland (1996, see also Sotaridona and Meijer, 2002) suggested that $p^*_{c'}$ be approximated using a piecewise linear function:

$$\hat{p}_r = \begin{cases} a + bQ_r & \text{if } 0 < Q_r \le 0.3 \\ [a + 0.3b] + 0.4b[Q_r - 0.3] & \text{if } 0.3 < Q_r \le 1 \end{cases} \tag{38}$$

where:

$r = 1,2,..., R$. R is the total number of groups each of which contains test-takers with the same number of incorrect answers.

Q_r = the percentage incorrect score of all test-takers with *r* incorrect answers

a, b = intercept and slope parameters

Sotaridona and Meijer (2002, see also Wollack, 2006) modified the K^* index further to produce the $\overline{K}_2$ index:

$$\overline{K}_2 = \sum_{g=m_{c'c}}^{w_s} \binom{w_s}{g} p^{*g}_2 (1 - p^*_2)^{w_s - g} \tag{39}$$

which uses a different approach to calculate the probability: $p^*_2 = b_0 + b_1 Q_r + b_2 Q_r^{\,2}$ where b_0, b_1 and b_2 are regression coefficients. Wollack (2006) suggested that the main advantage of $\overline{K}_2$ over the K indices is that the former uses information from all test-takers to compute the probability while the latter uses information only from those with the same number correct score as the source.

The g_2 index

The g_2 index index proposed by Frary et al. (1977) compares the number of identically answered items by the copier and the source against the expected number of identically answered items (see Wollack, 1997). If the answers from the source are treated as fixed and the probability of the copier answers item *j*, $P_C(u_{jS})$, exactly as the source's answer, u_{jS}, is known, then the expected number of items

that C answered identically as the source s, $E(h_{CS} \mid U_S)$, is the sum of the probabilities overall items J in the test:

$$E(h_{CS} \mid U_S) = \sum_{j=1}^{J} P_C(u_{iS})$$

where h_{CS} is the observed number of identically answered items, and U_S is the response vector for the source S. The variance of the number of matched answers is given by:

$$\sigma^2_{h_{CS}|U_S} = \sum_{j=1}^{J} P_C(u_{iS})[1 - P_C(u_{iS})]$$

The g_2 index for the pair is defined as:

$$g_2 = \frac{h_{CS} - \sum_{j=1}^{J} P_C(u_{iS})}{\sqrt{\sum_{j=1}^{J} P_C(u_{iS})[1 - P_C(u_{iS})]}} \qquad (40)$$

The statistic follows approximately the standard normal distribution. Large values of g_2 would indicate answer copying. The probabilities of C selecting each alternative of an item can be estimated by considering item difficulty and distractor difficulties and the ratio of the copier's raw score to the mean score of all test takers (see Wollack, 1997).

The ω index

The ω index proposed by Wollack (1997) is similar to the g_2 index discussed above. However, while g_2 is based on CTT, ω is based on IRT. Wollack (1997) suggested that since CTT item statistics are dependent on the trait levels of the test-takers included in the analysis, measures of the expected degree of similarity between a pair of examinees depend largely on the performance of the other test-takers on the test rather than only the two test-takers of interest. Under item response theory, the probability of a test-taker answering a specific item correctly is determined by his or her trait level and the characteristics of the item but independent of the other test-takers who take the test. Wollack also suggested that the various IRT-based person-fit statistics (such as those discussed before) used to identify aberrant response patterns do not depend on the similarity between the suspected copier's responses and those of a neighbouring test-taker and were therefore found to detect answer copying poorly. The ω index was developed specifically to detect answer copying under the IRT framework.

Wollack indicated that in investigating answer copying, the concern is not only with whether a pair of test-takers jointly answers an item correctly or incorrectly but also whether the same answer alternative was selected. This makes the IRT models for dichotomous items inappropriate for answer copying analysis. He used the nominal response model (NRM) developed by Bock (1972) to describe the interaction between the test-taker and an item. Under NRM, the probability, $P_{jk}(\theta_i)$, of test-taker i with ability θ_i select option k of an MCQ item j is given by the following equation:

$$P_{jk}(\theta_i) = \frac{\exp(\zeta_{jk} + \lambda_{jk}\theta_i)}{\sum_{l=1}^{m} \exp(\zeta_{jj} + \lambda_{jl}\theta_i)} \tag{41}$$

where:

m = number of alternatives
ζ_{jk} = item intercept
λ_{jk} = item slope

For each pair of test-takers for which copying is possible, the number of identically answered items (both correctly answered and incorrectly answered) can be counted which is denoted as h_{CS}. Treating the responses from the source as fixed and given the ability of the copier θ_C and the properties of the items in the test, the conditional probability that the copier selected the alternative k on item j which the source also selected is $P_{jk}(\theta_C)$ represented in Equation (40). The sum of the probabilities over all J items in the test will be the expected number of identical responses between the copier and the source, which is the expectation of h_{CS}:

$$E(h_{CS} \mid \theta_C, U_S, \xi) = \sum_{j=1}^{J} P_{jk}(\theta_C)$$

where:

U_S = the response vector of the source
ξ = the item parameter vectors

The variance, which is a measure of the variability, of the observed number of answer matches h_{CS} will be:

$$\text{var}(h_{CS}) = \sum_{j=1}^{J} P_{jk}(\theta_C)[1 - P_{jk}(\theta_C)]$$

The distribution of h_{CS} will approach normality as the number of items becomes sufficiently large. The ω index is defined as:

$$\omega = \frac{h_{CS} - E(h_{CS} \mid \theta_C, U_S, \xi)}{\sqrt{\text{var}(h_{CS})}} = \frac{h_{CS} - \sum_{j=1}^{J} P_{jk}(\theta_C)}{\sqrt{\sum_{j=1}^{J} P_{jk}(\theta_C)[1 - P_{jk}(\theta_C)]}} \quad (42)$$

ω becomes standard-normally distributed when the number of items is infinity. Values of ω can therefore be used to evaluate for statistical significance. The large the value of ω, the more likely the similarity in responses between the two test-takers resulted from answer copying.

The S_1 and S_2 indices

The S_1 index proposed by Sotaridona and Meijer (2003) is similar to the $\overline{K}_2$ index conceptually (see Wollack, 2006). However, S_1 uses the Poisson distribution to model the probability of match on an incorrect answer between the copier and the source. It also uses a log-linear model to estimate the probability parameter in the Poisson distribution for each group: $\log(\mu_r) = \beta_0 + \beta_1 w_r$, where w_r is the number incorrect score for test-takers in group *r*. Given the number incorrect answer of the source w_S, if the number of matched or identical incorrect answers between the copier c and the source c is w_{CS}, the probability of the matched incorrect answers that are greater than or equal to w_{CS} is the S_1 index for the pair:

$$S_1 = \sum_{g=w_{CS}}^{w_S} \frac{e^{\hat{\mu}} \hat{\mu}^g}{g!} \quad (43)$$

Small values of S_1 would suggest answer copying.

Sotaridona and Meijer (2003) extended the S_1 index to incorporate information on matched or identical correct answers between the copier and the source into a new copying index, the S_2 index. They argued that excluding the number of matched correct answers in the analysis of copying assumes that the copier knows all the correct answers to items both the copier and the source answered correctly, which may not always be true. They suggested that a test-taker may get a correct answer by copying or guessing. The K indices and the S_1 index are insensitive to copiers who only copy the correct answers and S_2 overcomes this limitation. For the S_1 index, the number of matched incorrect answers w_{CS} between the two test-takers is

used in the calculation. The formulae used to calculate the S_2 index is the same as that used for calculating the S_1 index. However, both the matched incorrect answers and the matched correct answers are used. The statistic used is m_{CS} which is calculated as the sum of the matched number incorrect answers w_{CS} and the matched number correct answers weighted by the likelihood of the match (see also Wollack, 2006). For a test-taker j_r in group *r* (with *r* number of incorrect answers), m_{CS} is estimated from the following equation

$$m_{CS} = w_{CS} + \sum_i \delta_{ij_r} = w_{CS} + \sum_i (d_1)\, e^{d_2 P_{ij_r}} I_{(u_{iS}=u_{i^*})}$$

where i^* denotes correct answers for item i, $I_{(u_{iS}=u_{i^*})}$ is an indicator function which equals 1 if the source *S* answered item *i* correctly and 0 if incorrectly, and P_{ij_r} is the percentage of test-takers in group *r* who match *S* on a correctly answered item *i*:

$$P_{ij_r} = \sum_{j=1}^{J_r} \frac{I_{(u_{iS}=u_{i^*})} I_{(u_{ij}=u_{iS})}}{J_r}$$

where $I_{(u_{ij}=u_{iS})}$ is also an indicator function which equals 1 if test-taker *j* and the source S answered identically to item *i* and 0 otherwise, and J_r is the total number of test-takers in group *r*. d_1 and d_2 are calculated from:

$$d_1 = \left(\frac{k+1}{k-1}\right)^{d_2 P_{ij_r}} I_{(u_{iS}=u_{i^*})}$$
$$d_2 = -(k+1)$$

where *k* is the number of item categories. The S_2 index is defined as:

$$S_2 = \sum_{g=m_{CS}}^{J} \frac{e^{\hat{\mu}} \hat{\mu}^g}{g!} \qquad (44)$$

S_2 follows a Poison distribution for which the parameter μ is estimated using the same loglinear model as that used for S_1. Small values of S_2 would suggest answer copying.

The Variable Match-Indices (VMIs) ξ and ξ^*

Belov (2011) proposed two indices, the Variable Match indices, which can be used to detect a variety of answer copying, including blind copying where two test-takers provide the same responses to different items that are in the same positions and

shift copying where the copier produce a response string to a set of items which is the same as the response string from the source but the positions of the items are different between the copier and the source (also see Bliss, 2012). These indices can be used in situations where the test is composed of two parts, an operational part which contains the same questions for all test takers and is used to generate test scores, and a variable part which may contain different items for different test takers. The test-takers will not know which part is operational and which part is variable. It is the responses to the items in the variable section that are used to investigate potential answer copying. A match happens when the potential copier answers incorrectly to item i of the copier's variable part and the potential source selected the same answer to $i+j$ of the source's variable part.

For two test-takers c, the potential copier, and s, the potential source, taking a linear test that is divided into two parts, the operational part (T) and the variable part (V), define two random variables:

$$\eta_{i,j} \equiv n_{i,j}(w_c, w_s) = \begin{cases} 1, & \text{c's incorrect answer to } i \in V_c \text{ and s' same answer to } i+j \in V_s \\ 0, & \text{otherewise} \end{cases}$$

$$\xi_j \equiv \xi_j(w_c, w_s) = \sum_{i \in V_c} \eta_{i,j}$$

where:

w_c = number of incorrect responses of c to the operational part T

w_s = number of incorrect responses of s to the operational part T

V_c = collection of items in the variable part of c

V_s = collection of items in the variable part of s

The VM-index ξ, which is conditioned on w_c and w_s, is defined as:

$$\xi \equiv \xi(w_c, w_s, l, u) = \sum_{j=1}^{u} \xi_j(w_c, w_s) \qquad (45)$$

where the summation parameters $l \le u$ make the VM-index sensitive to different types of coping:

When $l = u = 0$, ξ is sensitive to a blind-copy event.

When $l \le u < 0$, ξ is sensitive to a negative shift-copy event.

When $0 < l \le u$, ξ is sensitive to a positive shift-copy event.

When $l < 0 < u$, ξ is sensitive to all of the above events.

Belov used Monte Carlo method to estimate the critical values of the empirical distribution of ξ for a given significance level α.

Belov introduced an extension of the *VM*-Index, *VM*-Index* ξ^*, which is more conservative than the *VM*-Index (see also Bliss, 2012). This index uses the following two random variables:

$$\eta^*_{i,j} \equiv n^*_{i,j}(w_c, w_s) = \begin{cases} 1, & \text{c's incorrect answer to } i \in V_c \text{ and s's same but correct answer} \\ & \text{to } i + j \in V_s \\ 0, & \text{otherewise} \end{cases}$$

The index ξ^* is defined as:

$$\xi^* \equiv \xi^*(w_c, w_s, l, u) = \sum_{j=1}^{u} \sum_{i \in V_c} \eta^*_{i,j} \qquad (46)$$

Wesolowsky's Z similarity index

Wesolowsky (2000a) proposed a copying index which represents a modified version of the g_2 index proposed by Frary et al. (1977) and the ω index proposed by Wollack (1997) discussed above. Calculation of this index involves:

- work out the number of identical answered items in the test between two test-takers
- estimate the probability of a test-taker answering an item correctly
- estimate the probability distribution of the number of identically answered items by the two test-takers

The probability of test-taker *i* who answered item *j* correctly, $\hat{p}_{ij}$, is estimated from the following equation:

$$\hat{p}_{ij} = [1 - (1 - r_1)^{a_i}]^{1/a_i}$$

The parameter a_i is found by solving the following equation:

$$\frac{\sum_{j=1}^{J} \hat{p}_{ij}}{J} = c_i$$

where J is the number of questions in the test, and c_i is the proportion of questions answered correctly by test-taker *i.*

Given the observed number of matched items between two test-takers i and k, M_{ik}, the expected value, $\hat{\mu}_{ik}$, can be estimated from the probabilities that the two test-takers can answer each item in the test correctly or select specific wrong options:

$$\hat{\mu}_{ik} = \sum_{j=1}^{q} h_{ikj} = \sum_{j=1}^{q} [\hat{p}_{ij}\hat{p}_{jk} + (1-\hat{p}_{ij})(1-\hat{p}_{jk}) \sum_{t=1}^{v_i} \hat{w}_{tj}^2$$

where $\hat{w}_{tj}$ is the probability that, given the answer is wrong, wrong choice t is chosen on item j, and v_j is the number of wrong choices (distractors) of the item. The variance of the distribution of matched number of answers is estimated from:

$$\hat{\sigma}_{ik}^2 = \sum_{i=1}^{J} h_{ikj}(1-h_{ikj})$$

The Z index for the pair (i and k) is calculated from:

$$Z_{ik} = \frac{M_{ik} - 1/2 - \hat{\mu}_{ik}}{\hat{\sigma}_{ik}} \qquad (47)$$

Z_{jk} follows the standard normal distribution, and large values would suggest answer copying between the two test-takers. Compared with the g_2 and ω indices, this index also aims to reduce Type I error.

The M_4 similarity index

The similarity index M_4 proposed by Maynes (2005, see also Wollack and Maynes, 2011; Maynes, 2014a) decomposes the number of matching answers between two test-takers into two parts, with one related to the number of identical correct answers and the other the number of identical incorrect answers. It uses a generalised trinomial distribution to derive the exact distribution of the number of identical correct and incorrect answers.

The probability for a test-taker to select a particular answer alternative for an item depends on his/her ability and the item characteristics of the item and is modelled using the Bock' (1972) nominal response model. Under the assumption that two test-takers, j with ability θ_j and s with ability θ_j, work independently of each other when answering the item, the joint probability that j select a and s select a' on the same item i is given by the product of the probabilities $\pi_{ji_a}(\theta_j)$ and $\pi_{si_{a'}}(\theta_s)$ of them selecting their answers independently:

$$\pi(r_{ji} = a, r_{si} = a' \mid \theta_j, \theta_s) = \pi_{jsi} = \pi_{ji_a}(\theta_j)\pi_{si_{a'}}(\theta_s)$$

The probabilities for the two test-takers jointly to select the correct answer P_{ijs}, the identical incorrect alternative Q_{ijs}, and different alternatives R_{ijs} are:

$$P_{ijs} = \hat{\pi}_{ji_a}\hat{\pi}_{si_a} I(a = r_k)$$

$$Q_{ijs} = \sum_{a=1}^{A} \hat{\pi}_{ji_a}\hat{\pi}_{si_a} I(a \neq r_k)$$

$$R_{ijs} = 1 - P_{ijs} - Q_{ijs} = \sum_{a=1}^{A}\sum_{a'=1}^{A} \hat{\pi}_{ji_a}\hat{\pi}_{si_{a'}} I(a \neq a')$$

where r_k denotes the correct alternative (the key), $I(\cdot)$ is an indicator function which equals 1 if the statement in the parentheses is true and 0 otherwise, and *A* is the number of alternatives. The probability $f_{t,js}(m,n)$ that the two test-takers have *m* matching correct answers and *n* matching incorrect answers on *t* items in the test can be found using a recursion approach:

$$M_{4,js} = f_{t,js}(m,n) = P_{tjs} f_{t-1,js}(m-1,n) + Q_{tjs} f_{t-1,js}(m,n-1) + R_{tjs} f_{t-1,js}(m,n) \quad (48)$$

subject to the boundary conditions that $f_{1,js}(0,0) = 1$ when $m = n = 0$ and $f_{1,js}(0,0) = 0$ otherwise. Because the assumption on which the calculation of the statistic is based is that the two test-takers answered the two questions independently, small probability of the matched correct and incorrect answers would represent unlikely rare event. When $M_{4,js} = f_{t,js}(m,n)$ is less than 0.05, one could conclude that the probability of such a match by chance is small and therefore some kind of collusion between the test-takers is indicated. To control Type I error, the researchers suggested that M_4 is corrected by a multiplication factor of $(N-1)/2$ where N is the total number of test-takers. Pairs with $M_4 < 0.05$ are flagged.

3.2.2 Performance of copying and similarity indices

Wollack (1997) compared the performance of the ω and g_2 indices in detecting answer copying under different simulation conditions in terms of type of copying, percentage of items copied, proportion of test-takers engaged in copying and test length. Three types of copying was considered: random copying, difficulty-weighted copying and random string copying. His study found that ω was considerably better at controlling Type I error than g_2, with the Type I error rates for ω generally below the specified nominal Type I error rates while that for g_2 substantially inflated (above the nominal values) under the simulation conditions. It was found that the power of ω increases with test length and percentage of items copied and is insensitive to the type of copying. For a test of 80 items, the study found that the detection rate of ω was about 58% at $\alpha = 0.05$ when the proportion of items copied was 20%.

Wollack (2006) also used simulations involving manipulating responses from a real test to investigate the Type I error rate and detection power of eight copying indices. These included B, H, ω, S_1, S_2 and $\overline{K}_2$, and pairs of these indices. The types of copying considered in this study included random copying, string copying and mixed copying. Results from this study indicated that for the majority of these indices and their pairs, the Type I error rates were smaller than or similar to the nominal α levels (0.0005, 0.001, 0.005 and 0.01 respectively). The detection power of the indices varied substantially from 0% to over 90%, depending on the Type I error rates, percentages of items copied, test length and type of copying. It was found that for most of the simulation conditions and copying types, ω and the $\omega - H^*$ (H^* is a revised *H* index with critical values derived empirically) paired index out-performed the other indices, with ω being particularly powerful in detecting random copying and $\omega - H^*$ in detecting strings copying. Studies carried out by Zopluoglu (2016a) indicated that ω, the K-indices, S_1 and S_2 performed similarly for the datasets he analysed.

Belov (2011) compared the performance of three statistics, the K-index (K) and the Variable Match-Indices (ξ and ξ^*) in detecting blind copying and shift copying. His study indicated that, when slightly over 20% of the test-takers were involved in copying, the Type I error rates for these indices were generally below the nominal α values. When the proportion of test-takers involved in copying was small, ξ and ξ^* had better control of the Type I error rates. The detection rates of these indices varied, depending on the proportion of items copied. ξ and ξ^* generally had higher detection rates than K at all α values. In the case of blind or random copying, the detection rate was slightly below 70% for K, over 90% for ξ and slightly over 80% for ξ^* at $\alpha = 0.05$ when the proportion of items copied was 30%.

Simulation studies carried out by Wesolowsky (2000a) and Maynes (2014) indicated the two similarity indices Z_{ik} and M_4 generally had a Type I error below the nominal α level.

3.2.3 Practical applications and computer software

Many of the answer copying indices are sensitive to a range of factors involved in copying such as the nature of copying, the proportions of items being copied, test length, the proportions of test takers involved in copying. For practical applications, Wollack (2006) suggested that a straightforward approach is needed to select the index as the type and extent of copying is normally not known. An approach that possesses a good power and small Type I error rates for a range of possible copying conditions would be preferred.

Wollack (2006) suggested that many of the existing available indexes lack sufficient power at small α levels or when the amount of copying is relatively small. His work explored the utility of using multiple copying indexes in tandem to detect different types and amounts of answer copying. The results of his study suggest that using the ω and the revised H index (H^*) indices together may help improve power in these two areas without compromising the power in other conditions where power is already adequate. He concluded that ω is the index of choice to detect random copying and suggested that $\omega - H^*$ is generally most powerful to detect strings copying. If the items copied were small, H^* was powerful in detecting copying. He suggested that ω and $\omega - H^*$ would be good options for most applications. He also indicated that although H^* may be the best index in certain situations such as those involving, to a certain extent, strings copying, it however appears to be a poor choice when copying is dominated by random copying.

Zopluoglu (2016b) has developed an R packages that can be used to estimate a range of copying indices from test response data for tests composed of multiple choice questions. These include:

- the ω index
- K index and its variants
- the S_1 and S_2 indices
- the generalised binominal test

Wesolowsky (2000b) also developed the computer software SCheck to calculate the similarity index Z_{ik} he proposed.

3.3 Analysis based on comparison of performances on two subsets of items in the test

In situations where a test is constructed into sections and a particular section (or sections) may be prone to malpractice, the relationship between scores on different sections may be used to identify potential aberrant test-takers. This is different from person-fit statistics or copying indices discussed previously where the responses to individual items are examined for inconsistency. These indices may be used to detect aberrant responses associated with a range of behaviours such as answer copying, answer changes, item pre-knowledge, and inappropriate scoring.

Differential person functioning analysis

Smith and Davis-Becker (2011) proposed the use of differential person functioning (DPF) analysis for detecting cheating associated with prior knowledge of a proportion of the items in a test. DPF occurs when there are interactions between individual test-takers and classification of items in the test. DPF analysis is a way of examining the response behaviours of test-takers. The existence of DIF is a violation of

measurement invariance which constitutes an important aspect of validity. The fundamental assumption that DPF analysis can be used to identify cheating person is that a person's ability measure estimated using one set of items in the test should be similar to that estimated using the other set of items. If the ability estimate from the set of items prone to cheating is higher than that from the other set of item significantly, then the person has potentially behaved inappropriately. Smith and Davis-Becker conducted DPF analysis using the Rasch model, which involves:

- the items in the test is partitioned into two sets, with one set assumed to be prone to prior knowledge or other inappropriate behaviour and the other set not.
- conduct DPF analysis. There are different ways to do DPF analysis. One way is to analyse all person and item together to estimate item and person parameters. Once the item parameters are estimated, fix their values and re-estimate the ability of individual person on the two set of items. If the difference between the two ability estimates is large (greater than 0.6 logits) and is significant, then the person is flagged out as potentially inappropriately behaved.

Smith and Davis-Becker also investigated the sensitivity and stability of using this approach to detect potential cheaters with regard to the number of items not prone to inappropriate behaviour and the probabilities used to flag individuals. They found that with eight security items, a DPF contrast greater than 3, and flagging probabilities less than .005, the approach would result in 91% decision consistency, 1.1% Type I error rate, and a 7.9% Type II error rate.

The Kullback-Leibler Divergence (KLD) index

Belov et al. (2007) and Belov and Armstrong (2010) proposed to use the Kullback-Leibler Divergence (KLD) index to identify individual aberrant test-takers. Assuming that a test can be divided into two non-overlapping parts (*R* and *S*) or two sets of items. For a test-taker *e*, his/her posterior distributions of ability can be estimated separately based on responses to the two sets of items. These are denoted as $R(\theta_e)$ and $S(\theta_e)$ respectively. The KLD between the two distributions of ability is computed from the following equation:

$$KLD = \int_{-\infty}^{+\infty} R(\theta_e) \log \frac{R(\theta_e)}{S(\theta_e)} d\theta_e \qquad (49)$$

The distributions of ability is estimated using an IRT model. Relatively large values for KLD indicate significant difference in the test-taker's performance between the two parts. Belov and Armstrong (2010) described the following procedure for calculating KLD numerically:

- for each individual test-taker, construct his/her response vectors on the two parts: $r(r_1, r_1, ..., r_m)$ and $s(s_1, s_1, ..., s_n)$ where *m* and *n* are the total number of items in the

two parts respectively. The elements of the response vectors are either 1 or 0 (dichotomous items)

- to facilitate numerical approximation, the ability range is set to [-4,4] which is divided into h-1 intervals with h ability values $\{\theta_1, \theta_1, ..., \theta_h\}$ (h was set to 27 in their study)
- Bayesian posteriors for the two parts are computed based on the response vectors. The probabilities for the two parts can be approximated using the following equations:

$$R(\theta_l) = \frac{\prod_{j=1}^{m} P(r_j \mid \theta_l)}{\sum_{k=1}^{h} \prod_{j=1}^{m} P(r_j \mid \theta_k)}, \qquad l = 1,2,..., h$$

$$S(\theta_l) = \frac{\prod_{j=1}^{n} P(s_j \mid \theta_l)}{\sum_{k=1}^{h} \prod_{j=1}^{n} P(s_j \mid \theta_k)}, \qquad l = 1,2,..., h$$

where $P(r_j \mid \theta_k)$ and $P(s_j \mid \theta_k)$ are the probabilities corresponding to the responses given the ability level θ_k calculated using the IRT model employed

- the Kullback-Leibler divergence index is calculated from the following equation:

$$KLD = \sum_{l=1}^{h} R(\theta_l) \log \frac{R(\theta_l)}{S(\theta_l)} \qquad (50)$$

The value of KLD provides a measure of similarity between the two ability distributions. Large divergence values indicate a significant change in performance between R and S. KLD can also be used as a person-fit statistic. To be able to use KLD to flag aberrant respondent, a critical value corresponding to a desired level of significance α is needed. Such a critical value can be derived by producing the distribution of KLD using all test-takers. KLD will have high power to detect low-ability aberrant respondents.

Belov and Armstrong (2010) used KLD in combination with the answer copying K-index to detect answer copying. They found that KLD had better control of Type I error than the K-index. Belov (2014, 2015) also used KLD to identify aberrant behaviour associated with answer changes.

The Matched Percentile Index (MPI) and the Irregularity Index (IRI)

Based on the concept from equipercentile concordance used in test equating (Kolen and Brennan, 2008), Li et al. (2014) proposed the use of the matched percentile index (MPI) based on comparing the performances on parts of the test to identify test-takers with aberrant responses. In their study of a test composed of both MCQ

questions and construct response (CR) questions, the scores on the MCQ questions and those on the CR questions are equated to derive the MPI, which involved the following steps (see Li et al., 2014):

1. Convert observed MC scores to CR scores by identifying scores observed on the MC section that have the same percentile ranks as scores observed on the CR section. Compute the standard error of measurement (SEM) and obtain the error bands for the converted CR scores for a desired α level (for example, ± 3 SEM may be desired). Test-takers with converted score outside the error bands can be identified.
2. Similarly, convert the observed CR scores to MC scores by identifying scores observed on the CR section that have the same percentile ranks as scores observed on the MC section. The SEM is then computed to obtain the error bands for the converted MC scores for a desired α level. Test-takers with converted score outside the error bands can be identified.
3. Test-takers who are identified as outliers by both 1 and 2 above are flagged and assigned a value of True to the MPI. These test-takers are suspicious of test misconduct at the desired α level.

Li et al. also proposed to use the score on one part of the test (eg the MCQ section) to predict the score on the other part (eg the CR section). The predicted score then is compared with the observed score to derive the Irregularity Index (IRI). If the probability of the difference between the observed score and the predicted score is larger than chance, then the test-taker is flagged as an aberrant respondent. The following steps are needed to estimate IRI (see Li et al., 2014):

1. Based on the observed MCQ score of a test-taker, estimate his/her ability θ_{MC} Calculate the expected CR score and the standard error of estimation (SEE) for the test-taker using the ability estimate θ_{MC} from the MCQ scores and an IRT model for polytomous items.
2. IRI for the test-taker is calculated as the difference between the observed and expected CR scores divided by the estimated standard error of estimation.
3. For a desired α level, test-takers with IRI values above (or below) the critical value are flagged as aberrant respondents.

Simulation studies suggested that the false positive rates were less than 2.5% for the two indices and the detection power varied from negligible to over 60%, depending on simulation conditions.

The simple linear regression approach

Li et al. (2014) also used the simple linear regression method to look at the relationship between the difference scores between the MCQ section and CR section in the test, Y , and the ability estimates based on scores on the MCQ section θ_{MC} to identify potential aberrant respondents:

$$Y_i = a + b\theta_{MC,i} + \varepsilon_i \quad (51)$$

where i denotes the ith student, a and b are the model parameters, and ε_i is the residual. Test-takers with the observed difference score between the MCQ section and the CR section outside the 95% interval (or any other specified α level) of the predicted value are identified as potential aberrant respondents. If the residuals are standardised, test takers with the standardised residuals which are above (or below) 2 standard deviations (or any other desired values) are flagged.

Results from simulation studies showed that the power of these methods in identifying person with manipulated responses varied from nearly 7% to slightly over 35%, depending again on the type of manipulation simulated.

The Z-test statistic for difference scores

The difference between two sets of scores or ability estimates under the IRT framework from two sets of items in the test has also been used to identify aberrant respondents using a Z - test (see Guo and Drasgow, 2010; Li et al., 2014; Maynes, 2014b). If the two test scores or ability estimates are assumed to be measures of the proficiency of the test-taker, a significant difference between the two estimates or scores at the specified α level would suggest inconsistent performance on the two sets of items. If it is assumed that the ability estimate for a test-taker based on the first set of items is $\hat{\theta}_1$ with a standard error of estimation of se_1 and that estimated based on the second set of items is $\hat{\theta}_2$ with a standard error of estimation of se_2, the Z - test statistic can be calculated as:

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{se_1^2 + se_2^2}} \quad (52)$$

If the Z -statistic can be assumed to be normally distributed, for a desired α level, the corresponding critical value can be used for identifying aberrant test-takers. Results from simulation studies by Guo and Drasgow suggested that the Type I error rate was close to the nominal α value of 0.01 used, and the detection power varied from negligible to over 95%, depending on the difference in abilities estimated based on the two sets of items.

Use of the conditional probability distribution of score difference

Maynes (2014b) proposed an IRT approach to evaluate the differences of scores on two sets of items in a test to detect aberrant respondents, which overcomes the issue with estimating the standard error of score differences and the normality of Z score. The idea of this approach is that, for a test-taker, given his/her ability estimated using an IRT model, the conditional probability of the difference between

the two sum scores on the two sets of items can be computed. For the observed sum scores, if this probability is less than the critical value for a desired α level, the performances on the two sets of items are significantly different and the test-taker is flagged. The procedure involves the following steps:

- given the ability estimate θ of the test taker, the probability of score w on $k+1$ items $T_{k+1}(w|\theta)$ is calculated using a recurrence relation:

$$T_{k+1}(w|\theta) = \sum_s p_{k+1}(s|\theta)T_k(w-s), \quad T_0(0)=1 \text{ and } T_0(w)=0 \ \forall w \neq 0$$

  where $p_{k+1}(s|\theta)$ is the probability of scoring s on item $k+1$

- the joint probability of scoring x on the first set of items and y on the second set of items can be calculated from:

$$f(x,y|\theta) = T_x(x|\theta)T_y(y|\theta)$$

- given the observed score $S = x + y$ on the overall test, the conditional probability distribution of the difference score $d = x - y$ can be calculated from:

$$f(d|\theta,S) = \frac{T_x(x|\theta)T_y(S-x|\theta)}{f(S|\theta)} = \frac{T_x(\frac{S-d}{2}|\theta)T_y(\frac{S+d}{2}|\theta)}{f(S|\theta)} \qquad (53)$$

  Where the denominator is defined as:

$$f(S|\theta) = \sum_d T_x(\frac{S-d}{2}|\theta)T_y(\frac{S+d}{2}|\theta)$$

- for a given value of θ, total score S, score difference d and the desired α level level, if the conditional probability $f(d|\theta,S)$ is less than the critical value, the test-taker is flagged as aberrant.

Maynes applied this method to identify aberrant respondents associated with guessing, collusion or answer copying and performance on anchor items and unique items.

3.4 IRT models embedding aberrant behaviours

Item response theory models have also been proposed to model aberrant responses. The two models discussed here take into consideration item pre-knowledge and answer changing.

The Deterministic, Gated Item Response Theory Model (DGM) for item pre-knowledge

Shu et al (2013) proposed a model, the Deterministic, Gated Item Response Theory Model (DGM), that can be used to detect cheating resulting from item over-exposure.

This model classifies test-takers into two groups, cheaters and non-cheaters, by conditioning on two mutually exclusive types of items, the exposed items (compromised items) and the secure items (or unexposed items). Exposed items can be identified based on empirical exposure rates. The secure items are newly released. Shu et al. suggested that exposure of items acts like a gate through which cheating becomes possible through item pre-knowledge, while the secure items are not prone to cheating. The DGM identifies potential cheaters by computing their score gain in the exposed items from their scores on the secure items.

In the DGM, the observed item performance by a test-taker is decomposed into either their true proficiency function or a response function due to cheating ability. The model can be applied to tests composed of dichotomous items and can be expressed as:

$$\begin{aligned} &P(U_{ij}=1\,|\,\theta_{tj},\theta_{cj},b_i,T_i,I_i) \\ &\quad = P(U_{ij}=1\,|\,\theta_{tj})^{1-T_j}[(1-I_i)P(U_{ij}=1\,|\,\theta_{tj})+I_iP(U_{ij}=1\,|\,\theta_{cj})]^{T_j} \end{aligned} \tag{54}$$

where:

$P(U_{ij}=1\,|\,\theta_{tj},\theta_{cj},b_i,T_i,I_i)$ is the probability of scoring U_{ij} on item *i* by test taker *j* with a true ability of θ_{tj} and cheating ability of θ_{cj}

b_i = difficulty of item *i*

$T_j = \begin{cases} 1, & \text{when } \theta_{tj} < \theta_{cj} \\ 0, & \text{when } \theta_{tj} > \theta_{cj} \end{cases}$, an indicator variable used to flag cheaters

When $T_j = 1$, *j* is identified as a cheater.

$I_i = \begin{cases} 1, & \text{exposed item} \\ 0, & \text{secure item} \end{cases}$, model input defining the status of the item

$P(U_{ij}=1\,|\,\theta_{tj}) = \exp(\theta_{tj}-b_i)/[1+\exp(\theta_{tj}-b_i)]$, the Rasch model

$P(U_{ij}=1\,|\,\theta_{cj}) = \exp(\theta_{cj}-b_i)/[1+\exp(\theta_{cj}-b_i)]$, the Rasch model

When $T_j = 0$, the test-takers' responses to all items in the test are based on their true ability θ_t. When $T_j = 1$, the test-takers' response to unexposed items will be based on their true ability θ_t but their responses to exposed items will be based on their cheating ability θ_c. Shu et al. used Markov Chain Monte Carlo (MCMC) method to estimate model parameters of the DGM. They indicated that the difference between the true value of θ_{cj} and the true value of θ_{tj} should be zero when test-taker *j* is not a cheater. If the difference between the true values of θ_{cj} and θ_{tj} is greater than 0 then, the test-taker is a cheater:

$$\theta_{cj} - \theta_{tj} \begin{cases} = 0, & \text{for non - cheaters} \\ > 0, & \text{for cheaters} \end{cases}$$

The estimated values and the associated errors can be tested for significance in difference between the two variables. Test-takers can also be classified as cheaters or non-cheaters by setting a cut point $P_c (0 < P_c < 1)$ for the average $\hat{T}_j$ of the posterior samples of the indicator variable T_j:

$$T_j = \begin{cases} 1, & \hat{T}_j \geq P_c \text{ for cheaters} \\ 0, & \hat{T}_j < P_c \text{ for non - cheaters} \end{cases}$$

A higher cut point would indicate a higher confidence that the estimate of the cheating ability is greater than the estimate of the true ability.

Results from simulation studies suggested that, the specificity of the model, the percentage of non-cheaters correctly identified, was about 96% for all simulation conditions considered and out-performed the l_z index when the proportion of cheaters in the sample was 70%. The detection power of the model, or sensitivity, was found to be influenced by factors such as proportion of items exposed, the effectiveness of cheating and proportion of cheaters. The DGM model had more power in detecting effective cheaters showing a high level of score gain (ie, a high level of pre-knowledge) than less effective cheaters. With a proportion of cheaters at 5%, the model was able to detect about 80 % of high-effective cheaters. The detection rate was only 48% of high-effective cheaters when the proportion of cheaters was 70%. It was found that the DGM model had higher detection power than l_z index under all simulation conditions.

Modelling answer changes

Linden and Jeon (2012) attempted to model the probabilities of changes made to answers of items. Although their model can be applied to both paper and pencile based tests and computed based tests, their study focused on erasures made to answer sheets. A statistic based on wrong to right (WTR) changes (erasure) was proposed to identify unusual changes or aberrant respondents. It was assumed that test-takers have enough time to answer all items and review their answers. Two different stages of the response process were proposed: the first stage is to produce the initial responses to the items; the second (final) stage is to review the answers and make changes to the initial answers. Changes can be grouped into three categories:

- the initial correct response from the first stage was replaced by an incorrect answer in the second stage (right to wrong – RTW change)

- the initial wrong answer was replaced by another wrong answer (wrong to wrong – WTW change)
- the initial wrong answer was replaced by a correct answer (wrong to right – WTR change). The statistic *E* below is the total number of WTR changes in the test

The following steps are involved to derive the critical value of the statistic:

- based on the responses from the first stage using the 3PL model (for MCQ test), the abilities of the test takers are estimated
- for the second stage, the abilities of the test takers are fixed as the values estimated from the first stage. For a WTR change in the second stage, the probability is modelled using a 2PL model as follows:

$$P_{ni} = \Pr\{U_{ni}^2 = 1 \mid U_{ni}^1 = 0\} = \frac{\exp[a_{0i}(\theta_n^1 - b_{0i})]}{1 + \exp[a_{0i}(\theta_n^1 - b_{0i})]} \quad (55)$$

where:

n = the nth test-taker
i = the ith item in the test

θ_n^1 = the ability of test taker n estimated using responses from the first stage

U_{ni}^2 and U_{ni}^1 = the second and first responses from test taker n on item i (1 for a correct and 0 for an incorrect response)

a_{0i} and b_{0i} = item parameters of item i

The item parameters $\{a_{0i}\}$ and $\{b_{0i}\}$ can be estimated using a subset of the final response data constructed by selecting responses being incorrect (ie $U_{ni}^1 = 0$) at the first stage. A logistic regression approach was used to estimate the item parameters.

- given the number of WTR changes E_n for test-taker n with known ability on J_n changed items with known item parameters, the probability distribution of possible number of WTR changes E on the J_n items are calculated by the generating function with the recursive method proposed by Lord and Wingersky (1984):

$$\Pr\{E = e\}, \quad e = 0,1,..., E_n \quad (56)$$

For a given level of significance α, identify e as the critical value for which the probability $\Pr\{E = e\}$ is less than α from the list of abilities listed from the above equation. The test-taker will be identified as an aberrant respondent if the total number of E_n is large or equal to the critical value: $E_n = e_n^*$.

When this model was applied to the responses of 2555 Grade 3 students to 65 mathematics items, 2.6% of the students were found to have aberrant answer changes at $\alpha = 0.05$ level.

4. Analysis of aberrant response patterns and unusual test scores for groups

While research studying aberrant responses at individual test-taker level started in the early 1920s, research investigating aberrant responses or anomalous scores at class or school levels only received increased attention over the last two decades. This partly reflects increasing report in the media of test collusion involving large number of individuals at different levels of the system (see Thiessen, 2008; Wollack and Maynes, 2011; Plackner and Primoli, 2014). A wide range of statistical techniques have been developed and used to identify groups (test centres, classes and schools) with anomalous test results. These generally involve the following analyses at group level:

- wrong-to-right (WTR) erasure rates
- test score and response patterns
- growth rates over time
- test score distributions
- relationships in performance between parts of the test
- relationships between test performance and other variables

4.1 Analysis based on wrong to right (WTR) answer changes

Analysis of wrong-to-right erasures has been used by testing companies or other authorities to identify class or school level cheating (see Wibowo et al., 2013; McClintock, 2015). Wibowo et al. (2013) described the conventional approach used to conduct erasure analysis: For a unit or group (class or school) u, the WTR erasure rate $\overline{M}_u$ is defined as the average number of erasures within the unit:

$$\overline{M}_u = \frac{1}{N_u} \sum_{k=1}^{N_u} M_{u,k}$$

Where:

$M_{u,k}$ = the number of erasures of student k in unit u

N_u = the number of students in unit u

Given the mean erasure rate μ and standard deviation σ for the population, the sampling distribution of samples with a size of N_u will be normally distributed with a

mean of μ and standard deviation of $\sigma/\sqrt{N_u}$. For the sample from a specific unit *u*, if the observed mean erasure rate is significantly higher than the population mean for a pre-specified level of significance, it is flagged as an outlier and may be subject to further investigation for potential test collusion. The significance level is specified as the number of standard deviation Δ that the unit mean departs from the population mean:

$$\overline{M}_u > \mu + \Delta\frac{\sigma}{\sqrt{N_u}}$$

Δ takes integers such as 3, 4 or 5. The conventional method is not appropriate for units with small size (eg <100) due to inflated Type I error rates.

Wibowo et al (2013) used a Poisson-gamma distribution to model the distribution *P* of the number of WTR erasures within a unit to flag units with unusual number of erasures:

$$P(M_k = m \mid r, p) = \frac{\Gamma(r+m)}{m!\Gamma(r)}(1-p)^r\, p^m$$

where:

$m = 0,1,...$

r, p = model parameters to be estimated and $0 \le p < 1$

For a unit with the total number of WTR erasures $S_u = \sum_{k=1}^{N_u} M_k$, the distribution of S_u will also follow a Poisson-gamma distribution with parameters rN_u and p:

$$P(S_u = s_u \mid r, p, N_u) = \frac{\Gamma(rN_u + s_u)}{s_u!\Gamma(rN_u)}(1-p)^{rN_u}\, p^{s_u} \qquad (57)$$

where $s_u = 0,1...$ is the value of the random variable S_u. For a specific unit *u* with a total number of WTR erasures s_u, if the probability is less than the specified significant level α, then the unit is flag as having unusual number of WTR erasures:

$$P(S_u \ge s_u \mid \hat{r}, \hat{p}, N_u) < \alpha \qquad (58)$$

Results from simulation studies suggested this method had better control of the Type I error rates than the traditional method. The Type I error rates were generally smaller or close to the nominal α values under the simulation conditions.

4.2 Analysis based on response patterns and test scores

4.2.1 Analysis based on similarity indices

Use of similarity indices in conjunction with nearest neighbour clustering approach

Wollack and Maynes (2011, 2017) present an approach which could be used to detect clusters of test-takers engaged in test collusion. The approach is based on analysis of the similarity of answers between test takers and does not require that the groups of potentially contaminated examinees be identified a priori and can be applied to data from a single test administration. The method can be used to identify individuals whose test scores are of questionable validity.

The method uses the nearest neighbour (or single linkage) clustering in conjunction with an answer similarity index used to characterise the degree of similarity in the answers between two test takers. The approach involves the following steps:

- computation of the answer similarity index and set the threshold for flagging pairs of test takers with unusual similarities in their answers. The answer similarity statistic used by Wollack and Maynes is the M_4 index proposed by Maynes (2005)
- once values of the similarity between all possible pairs of test takers have been calculated, aberrant respondents can be identified using the threshold. These respondents are then grouped into clusters using the nearest neighbour clustering method with their paired similarity data. In this clustering approach, all linked test-takers are grouped into one cluster. That is, two clusters S and T, which contain two sets of test-takers $N_s(s=1,2...,n_s)$ and $N_t(t=1,2...,n_t)$ respectively, are clustered together if the similarity index $S(s_i,t_j)$ for pair $[s_i,t_j]$ exceeds the pre-defined threshold for at least one $[s_i,t_j]$ pair between the two clusters

The researchers also used a statistical model to simulate the impact of collusion on the probability of selecting identical alternatives between two test-takers using simulated item response data. Their results indicated that it is possible to recover clusters of inter-related test-takers, provided the amount and magnitude of collusion is reasonably high. Cluster integrity, which is a measure of the extent to which the grouped clusters are interpretable, improves as the cluster effect and the number of exposed items increases. The Type I error rates were found generally to be below the nominal significance level at $\alpha = 0.05$ used in their study. The detection power was influenced primarily by the number of items compromised and the collusion strength.

Use of group average of similarity measures

The similarity indices such as the Z and M_4 used for flagging individuals with unusual response patterns discussed before can also be used to flag groups with unusual number of aberrant respondents. Sotaridona et al. (2014) presented a standardised non-parametric matching index $Z_{nn'}$ for flagging pairs of individuals taking an MCQ test composed of J items:

$$Z_{nn'} = \frac{M_{nn'} - \sum_{i=1}^{I} P_{nn',i}}{\sqrt{\sum_{i=1}^{I} P_{nn',i}(1 - P_{nn',i})}} \tag{59}$$

where:

$n, n'(n \neq n')$ is test taker pair (n, n')

$P_{nn',i} = \sum_{k=1}^{k_i} P_{i,n,k} P_{i,n',k}$ is the expected probability that (n, n') will match on their response to item i, k is response category, $P_{i,n,k}$ and $P_{i,n',k}$ are the response probabilities, and k_i is the number of response categories

$M_{nn'}$ is the number of matched items

$Z_{nn'}$ is asymptotically normally distributed and can be used to flag pairs with unusual matched number of items for a specified level of significance α. For class u with a total number of test-takers N_u and a test-taker n in u, the average of her/his $Z_{nn'}$ across the test takers in the unit can be calculated as:

$$\overline{Z}_n = \frac{\sum_{n'=1, n' \neq n}^{N_u} Z_{nn''}}{N_u - 1}$$

For this distribution of the average values, the mean μ_u and deviation σ_u can be calculated. The population mean μ of the above statistic and its standard deviation σ can be calculated. For a specific class with N_u test-takers, if the group mean μ_u is significantly greater than the population mean μ for a given level of significance α, the class is flagged as performed abnormally on the test:

$$T_u = \frac{\mu_u - \mu}{\sigma_u / \sqrt{N_u}} \tag{60}$$

T_u is assumed to be asymptotically normally distributed and can be used to flag suspicious classes. Sotaridona et al. (2014) subsequently improved the method by producing a parametric statistic using Bock's nominal response model to estimate matching probabilities between pairs of individuals. Real test data were manipulated

to test the power and the Type I error rate of the index. It was found that the Type I error rates were generally below or close to the nominal levels. For a given level of α, the detection rate varied with the proportion of items copied. At $\alpha = 0.02$, the detection rate was almost 100% when the proportion of items copied was over 40%. The parametric approach also out-performed the non-parametric approach.

4.2.2 Analysis based on person-fit statistics

Use of factor analysis for grouping aberrant respondents

Zhang et al. (2011) used Q-type factor analysis to cluster aberrant respondents identified using person-fit statistics further into different groups. Each group contains respondents with similar aberrant responses. Different groups may show aberrant responses on different set of items. Their approach involves the following main steps:

- select person-fit indices from the existing research literature. In their study, they used the unweighted U statistic for the Rasch model
- establish thresholds with simulated data for the chosen person-fit statistics and use them to identify test-takers with aberrant item responses
- assign aberrance scores to items for each test-taker flagged by person-fit indices to construct aberrant response vector. In their study, the aberrance response vector is constructed using the following procedure:

  1. The original response data from all test-takers is analysed using the Rasch model for dichotomous items, and the unweighted person fit statistic U is calculated for each person:

  $$U_n = \frac{1}{J}\sum_{j=1}^{J}\frac{(X_{nj} - P_{nj1})^2}{P_{nj1}P_{nj0}} = \frac{1}{J}\sum_{j=1}^{J}V_{nj} \qquad (61)$$

  V_{nj} is the variance of the score of person n on item j.
  2. Set the threshold for U_0 to identify persons with under-fit to the Rasch model (ie persons with variability in their item scores larger than the Rasch model predicted).
  3. For each item j in the test, the aberrant response score by person n is determined using the threshold U_0:

  $$Y_{nj} = \begin{cases} 1 & \text{if } V_{nj} > U_0 \\ 0 & \text{otherwise} \end{cases}$$

  The aberrance response vector is $Y_n = \{Y_{nj}\}$.

- Create the matrix of inner product of aberrant response vectors for the identified aberrant respondents. The researchers indicated that clustering can be based on

the Euclidean distance or dot product between two response vectors. However, they argued that while Euclidean distance is more likely to measure the dissimilarity between two aberrance response vectors, the dot product measures the similarity between the vectors:

$$d_{Euc}(Y_1, Y_2) = \sqrt{(Y_1 - Y_2)(Y_1 - Y_2)^T} = \sqrt{(y_{1j} - y_{2j})^2} \quad (62)$$

$$dot(Y_1, Y_2) = Y_1 \cdot Y_2 = \sum_j y_{1j} y_{2j} \quad (63)$$

They chose the dot product approach over the Euclidean distance as the purpose was to identify test-takers whose test responses are aberrant and answered in the similar ways. The dot product is however insensitive in distinguishing pairs of test-takers when they answered the same set of items correctly but had various patterns of inconsistent responses to other items. To overcome this issue, the aberrance scores on each item are standardised before producing the dot product matrix. The standardised aberrance response vectors can be expressed as a matrix, and the matrix of inner product can be generated as the correlation matrix between aberrant respondents which is different from conventional correlation matrix where correlations between variables are used.

- Analyse the matrix with factor analysis, the Q-type factor analysis, with rotation such as the Varmax rotation technique, and use factor loadings on individual factors to group test-takers with similar aberrant responses.

This approach intends to cluster aberrant respondents into groups which may possess shared pre-knowledge of test content. The detection power was affected by the number of compromised items and the number of test-takers with pre-knowledge. When 5% of the items (a total of 200) were compromised, the average detection rate was around 38%, higher than that identified using l_z or the Caution Index *C*. The detection rate was close to 100% when 20% of the 200 items used were compromised.

Detection of test collusion using Kullback-Leibler Divergence

Belov (2013, 2014, 2016) proposed the use of the Kullback-Leibler divergence index to investigate aberrant test performance, particularly test collusion which involves large scale sharing of test materials (including answers to test items) at test centre level. Here, the definition of test centre is not limited by the geographic location. His approach works in two stages:

- stage 1: test centres with an unusual distribution of a person-fit statistic are identified using a statistic related to Kullback–Leibler divergence. For a centre c belonging to the collection of all centres C_s ($c \in C_s$), this statistic g_c is defined as:

$$g_c = \sum_{x \in C} [D(H_c \parallel H_x) + D(H_x \parallel H_c)] \quad (64)$$

where:

H_c and H_x = the empirical distribution of the person-fit statistic used for centres *c* and centre *x*

$D(H_c \| H_x)$ is the Kullback-Leibler divergence defined for a finite set of K values $\{d_1, d_2, ..., d_k\}$ used to represent the distribution of H_c and H_x which can be calculated from the following equation:

$$D(H_c \| H_x) = \sum_{k=1}^{K} H_c(d_k) \ln \frac{H_c(d_k)}{H_x(d_k)} \quad (65)$$

g_c is a measure of dissimilarity between the distribution of the person-fit statistic for test centre *c* and the distributions for the other centres. The definition of the statistic g_c balances the asymmetry of the Kullback-Leibler divergence. $D(H_c \| H_c) = 0$.

- stage : test-takers from identified test centres are analysed further using the person-fit statistic, where the critical value of the fit statistic used to detect aberrant respondents is computed using data from non-aberrant centres only

Computer simulation studies were conducted to investigate the power of this approach for different conditions under which items are compromised and Type I error rates. The Type I error rates were below the nominal levels. The detection rates were over 90% at $\alpha = 0.05$ for the conditions simulated. The approach was found to be effective in computer adaptive testing for detecting groups of test-takers with item pre-knowledge (accessed one or more subsets of items prior to the exam). Belov suggested that this approach is extremely flexible as any existing person-fit statistic used to detect aberrant test-takers can be used. Further, this approach can be applied to many forms of testing, including paper-and-pencil testing, computer-based testing, multi-stage testing (MST), and computer adaptive testing.

Use of group proportion of persons identified as aberrant respondents by fit statistics

Although the various person-fit statistics discussed previously which are used to identify individual test-takers with aberrant responses, given the specified level of significance α, the proportion of persons identified as aberrant respondents in a group (eg a class or a school) can be calculated. This group proportion may then be compared with the proportion of aberrant respondents observed for the population. If the group proportion of aberrant respondents is statistically significantly higher than the population value, the group may be assumed to have performed abnormally in relation to other groups.

4.2.3 Analysis based on item responses and test score distributions within individual groups

Factor and cluster analysis based on item responses and test score distributions

When investigating score anomaly at class level associated with score manipulation in the Italian standardised national tests for primary and secondary schools, Quintano et al. (2009) used a fuzzy *k*-means clustering approach which is based on four class-level indicators of test scores to identify outlier classes and correct class scores. These assessments contain both closed-form and open-ended items. The four indicators used by the researchers are:

- Class mean score on the test P_s
- Standard deviation of scores of the class σ
- Class non-response rate R_{nr} which is defined as:

$$R_{nr} = \frac{\sum_{i=1}^{N_s} J_{nr,i}}{N_s J} \quad (66)$$

where:

N_s is the number of number of students in the class

J is the number of number of items in the test taken by the class

$J_{nr,i}$ is the number of items not responded by student i

- Homogeneity index of answers H_o which is defined as:

$$H_o = \frac{1}{J}\sum_{i=1}^{J}\left(1-\sum_{j=1}^{h_i}\left(\frac{n_{i,j}}{N_s}\right)^2\right) \quad (67)$$

where:

h_i is the number of alternative answers to item i

$n_{i,j}$ is the number of students in the class that gave the jth answer to item i

Further analysis of the dimensionality of the indicators was undertaken using exploratory factor analysis with principal component (PC) extraction in order to select a set of underlying factors for clustering analysis. For their study, the researchers found that the first two components accounted for over 90% of the total variance. The first component was highly correlated with all the four indicators, while the second component was highly correlated with class non-response rate indicator. The correlation between the first factor and the class mean score was highly negative, while the correlations with both the standard deviation and answer homogeneity

were highly positive. The second component was highly correlated with class non-response rate. The researchers suggested that the first component could be interpreted as the "outliers identification axis" and the second component the "index of class collaboration to survey".

The classes were then classified into 8 groups using a fuzzy version of the non-overlapping k-means clustering with a value of 2 for the fuzzy parameter r (see Bezdek, 1981) based on the two principal components identified. This involves minimizing the following objective function using the repetition method:

$$J_{FKM} = \sum_{n=1}^{N_c} \sum_{s=1}^{S_c} P_{ns}^{r} d_{ns}^{2} \tag{68}$$

where:

N_c = total number of classes

S_c = total number of clusters (8)

r = fuzziness parameter (=2. r=0 represents normal non-overlapping clustering)

d_{ns} = the distance between class n and the centroid of cluster s

$P_{ns} \in [0,1]$ = the cluster membership degree of class n belonging to cluster s and $\sum_{s=1}^{S_c} P_{ns} = 1$. P_{ns} is calculated from:

$$P_{ns} = \frac{1}{\sum_{t=1}^{S} (d_{ns} / d_{nt})^{2/(r-1)}}$$

The centroids of the clusters are then projected onto the two factor components to identify the outlying cluster. The researchers then used the class membership degree for the outlying cluster as a manipulation indicator to correct the class mean score. Based on comparison of performances between classes with external monitors and those without, Battistin et al. (2014, also see Angrist et al., 2014) further improved the estimation of the manipulation index.

Comparison of item responses and test score distributions between groups taking the same test under different conditions – a likelihood approach

In the Italian standardised national tests discussed above, the majority of the students take the tests in their own classrooms, invigilated by teachers from their own schools. However, these teachers are not currently teaching the classes they are invigilating. The teachers are also responsible for marking students' work where needed, transcribing the answers and sending the results back to the National Institute for the Evaluation of the Education System (INVALSI) for analysis. A

proportion of the classes (about 10%) is also randomly selected and invigilated by external monitors. These external monitors perform the same tasks as the school teachers but have no prior connection to the schools they are assigned to. Fernández (2016) uses this as a large-scale natural experiment in which classes invigilated by external monitors were treated as the treatment group while those by teachers as the control group and adopts a likelihood approach to detect potential test score manipulation in classes where school teachers were invigilators. This approach is based on the comparison of score distributions between the treatment and control groups after controlling for the effects of other factors which could also affect students' test scores. Classes with unlikely outcomes are identified through low values for the likelihood function of their score distribution. The likelihood values are also used to adjust class mean scores. The approach can overcome some of the limitations associated with the fuzzy *k*-means clustering approach used by Quintano et al. (2009) discussed above.

Steps involved in implementing this approach include:

- Model a response y_{icj} of student *i* in class *c* to item *j* in the test using latent variables:

  $$y_{icj} = 1(y^*_{icj} \geq 0)$$
  $$y^*_{icj} = \eta_{ic} + \xi_j + \varepsilon_{icj}$$

  where η_{ic}, ξ_j and ε_{icj} are the latent variables representing the individual-class effect, question effect and the individual-class-question iid shock respectively. The individual-class effect are treated as random and question effect as fixed
- after the individual-class effects are accounted for, the answers of two students are independent. This make it possible to construct a likelihood function of score distribution of the classes. The likelihood function is modelled separately for the treatment group and the control group
- for each class, a likelihood, $\hat{l}_c$, can be estimated for its score distribution
  Comparison of the probability density function of the likelihood between the treatment group and the control group can be used to detect score manipulation in classes in the control group. That is classes with unlikely results (or with small probabilities)
- the cumulative distribution functions (cdf) of the likelihood of the classes are constructed for the treatment group and the control group separately, which are denoted as $F_{L,TR}(l)$ and $F_{L,CO}(l)$ respectively
- the cumulative distribution functions are used to adjust the likelihood of the classes in the control group:

  $$\tilde{l}_c = \frac{F_{L,CO}(\hat{l}_c)}{F_{L,TR}(\hat{l}_c)} \hat{l}_c \qquad (69)$$

This will result in the adjusted cdf of the likelihood for the classes in the control group to be the same as that for the classes in the treatment group

- the adjusted and unadjusted likelihood values are used to adjust the class scores for classes in the control group

4.2.4 Analysis based on relationships between scores on subsets of items within the test

The simple linear regression approach

Li et al. (2014) used the simple linear regression method to look at the relationship between the difference scores between the multiple choice question (MCQ) section and constructed response (CR) section in the test, Y, and the ability estimates based on scores on the MCQ section θ_{MC} to identify groups of test-takers who might have performed on the test unusually. Their approach involves:

- estimate the ability of persons based on their responses on the MCQ section of the test
- or each test-taker, work out the difference score between the raw score on the MCQ section and the raw score on the CR section
- for a group, such as a class or a school, work out the mean of difference scores $\bar{Y}$ and the mean of the ability estimates $\bar{\theta}_{MC}$. The mean difference score is then regressed on the mean ability estimate:

$$\bar{Y}_i = a + b\bar{\theta}_{MC,i} + \varepsilon_i \qquad (70)$$

where i denotes the ith group, a and b are the model parameters, and ε_i is the residual. Groups with the observed mean difference score between the MCQ section and the CR section outside the 95% interval (or any other specified α level) of the predicted value are identified as abnormal groups. If the residuals are standardised, groups with the standardised residuals which are above (or below) 2 standard deviations (or any other desired values) are flagged. The detection rate at class level was found to be affected by the type of simulated irregularity.

4.3 Analysis based on similarity of response patterns and other variables over time

The Jacob and Levitt approach

To investigate whether there was test collusion in individual classes, Jacob and Levitt (2003) developed a method which uses two class level indices, with one related to the unexpected class score fluctuations in terms of score gains between two consecutive years and the other related to unusual similarity in item response patterns from the students in the same class for blocks of items. Classes which have

high values on both indices are flagged as potential instances of test collusion. Thiessen (2007, 2008) and Wollack and Maynes (2011) provided a summary of the method developed by Jacob and Levitt which is further summarised below.

Index for unusual test score fluctuations

If test scores from different years are placed on the same scale, it is normally expected that most of the students' scores increase at a relatively constant rate over time although variability in score gains between students exists as they are affected by a range of factors. For a specific class, if the majority of the students have large test score gains in one year but followed by small score gains (or loss) in the next year, then unexpected test score fluctuation has happened. Assuming that year *t* is the year in which the test of interest was administered, the unexpected test score fluctuation index SC_{cbt} is derived using the following procedure:

- work out the average test score gains from t-1 to t and from t to t+1 for all students in the class.
- work out the percentile rank of the class' average test score gains relative to all other classes in that same subject, grade, and year. The percentile ranks of growth are $rk\_gn_{c,b,t}$ from year t-1 to year t and $rk\_gn_{c,b,t+1}$ from t to t+1 respectively.
- the index SC_{cbt} is defined as:

$$SC_{cbt} = (rk\_gn_{c,b,t})^2 + (1 - rk\_gn_{c,b,t+1})^2 \qquad (71)$$

As is clear from the above definition, the index takes higher values for classes that show large score gains this year and small score gains next year. It is also clear that the use of squares in the definition gives more weight to large score gains this year and large score decline the following year. Classes that have values in the top 95th percentile of SC_{cbt} will be flagged as having unexpected test score fluctuations.

Index for unusual item response patterns within a class

The second index, ANS_{cbt}, is used to identify unexpected item response patterns in students' answers within the class. This index combines the following four measures:

- Measure for unlikely block of identical answers by students on consecutive questions.

  A multinomial logit model was used to predict the likelihood of each student choosing each possible answer on each question, taking into account the student's past test scores, future test scores, and background characteristics. The block of identical answers from the students in the class that were least likely to

have arisen by chance was identified by searching all combinations of students and consecutive questions.

If each student in the classroom has unique responses from item *m* to item *n*, then there will be a distinct value of this index for each student in the class. If all students in the classroom have identical responses across these items, then there will only be one value of this index (and the value will be extremely small).

The calculations are repeated for all strings from a length of 3 items to a length of 7 items.

It is to be noted that the values yielded by these calculations will be smaller as: (1) the number of students with identical responses increase, (2) the length of the string of identical responses increase. Thus, smaller values are associated with more improbable answer strings within a classroom. The minimum value of this measure for each classroom is recorded as measure 1:

- Measure for the correlation in student responses across the test

  This measure is intended to capture more general patterns of similarity in student responses beyond just identical blocks of answers. It is derived from the residuals for each item choice for each student in the class calculated based on the predicted category probabilities from the multinomial logit model. This is to a degree also a measure of within-class correlation in student responses. This measure, measure 2, will take high values for a class if students in the class tend to give the same answers to many questions.

- Measure for variability in correlation between questions

  As there can be many reasons to account for high values of within-class correlation represented by the second measure (for example, specific topic areas have received particular attention during the school year which could result in correct answers from students to the relevant questions), the third measure, measure 3, which measures the variance in the degree of correlation across questions and is calculated as the variance of question residuals from the second measure, is introduced to detect potential cheating. If the answers for multiple students on selected questions were changed, there would be high within-class correlation on those questions, while the within-class correlation on other un-changed questions would likely to be typical. This would lead to larger cross-question variance in correlations than normal in the cheating classes.

- Measure for unusual response patterns for students with the same test scores

  This measure compares the responses from students within a class to those from other students in the system who have obtained the same test scores. It is used to identify students who answered difficulty questions correctly but easy questions

incorrectly or missed the easy questions, which may be an indication of cheating. The measure is calculated based on comparing the class level item scores with the population item scores (conditioned on the same total test scores). Large values of this measure would suggest that the responses from a large number of students in the class deviated from those in the system who have similar total test scores.

Classes are ranked on each of the four measures discussed above. The percentile ranks are then squared and summed to form an overall measure for the second index:

$$ANS_{cbt} = (rk\_m1_{c,b,t})^2 + (rk\_m2_{c,b,t})^2 + (rk\_m3_{c,b,t})^2 + (rk\_m4_{c,b,t})^2 \qquad (72)$$

Classes with overall ranking on this index above the 95th percentile are identified as having unusual patterns in the responses from students.

Identifying cheating classes

Jacob and Levitt argued that taken individually, the above two indices do not detect teachers who manipulate the responses of their students. It is possible that some classes will have unexpected score fluctuations and some classes unusual response patterns. However, the likelihood of a class that scores high on both indicators should be small. For classes where no collusion took place, the two indices should not be highly correlated. In contrast, if a teacher manipulates students' responses, strong correlation between the two indices would be expected. For classes with the 95th percentile ranking on both indices, there is potential that test collusion has happened. Application of this method to real data suggested that about 4-5% of the classrooms studied potentially had cheated every year.

The two-proportion Z-score approach

Gaertner and McBride (2017) used the two-proportion *z*-score which is based on the difference in the school's pass rate between two years and the difference in the population pass rate between the two years to identify schools with anomalous changes in pass rates over time. The effect size of the difference in pass rates across the years is expressed as Cohen's h which is considered to be large if greater than 0.8. Their simulation study suggests that the z-score approach was effective when cheating occurred in a large school.

The multilevel logistic regression (MLR) approach

Gaertner and McBride (2017) also used a two-level logistic regression model to investigate school level change in pass rate over time. A student's likelihood of attaining a passing score in the second year is modelled as a function of the school she/he attends and the prior-year pass rate at that school. The school level residuals

are used to identify anomalous schools. Their simulation study indicates that the MLR approach was effective in detecting cheating at small schools.

The Bayesian hierarchical linear growth model approach

Skorupski and Egan (2011, 2014) and Skorupski et al. (2017) proposed an approach to detect cheating and aberrance at group level that uses a Bayesian hierarchical linear model (HLM) to describe growth in performance on state-wide assessments (SWAs) over time (three years in their studies). In the model, the scores from individuals over time are nested within students who in turn are nested within groups. The model can be expressed as follows:

$$Y_{igt} = \beta_0 + \beta_{1g}(G) + \beta_{2t}(T) + \beta_{3gt}(GT) + \varepsilon_{igt} \quad (73)$$

where:

Y_{igt} = vertically linked score of students i in group g with N_g students at time t

G, T and GT = group effect, time effect and group-time interaction effect respectively

β_0, β_{1g}, β_{2t} and β_{3gt} = intercept, main effect for group, main effect for time and interaction effect between group and time

ε_{igt} = random error with an expected value of zero

Unusually large group-time interaction would suggest potential aberrance. A statistic, delta (δ_{gt}), was proposed to evaluate the effect or size of group-time interaction:

$$\delta_{gt} = \frac{\beta_{3gt} - 0}{\sqrt{\sigma_t^2}} \quad (74)$$

A critical value of 0.5 was suggested for flagging aberrant groups. Results from simulation studies indicated that the detection rate varied from 55% to 83% with the Type I error rates varying from 1% to 2% for the simulated conditions.

4.4 Analysis based on relationship with other variables

This section discusses methods used to identify groups with anomalous test results that are based on relationships between the test being investigated and other variables. This type of analysis is particularly useful when item response vectors for individuals are not available. For example, results from school-based teacher assessment (SBTA) or non-exam assessment (NEA) normally just report an aggregated score at the overall assessment level. There is potential for inappropriate marking/scoring of students' work to take place in SBTAs and NEAs (eg Ofqual, 2012).

Use of the cumulative logit regression model

Clark et al. (2013, 2017) used the cumulative logit regression model to investigate groups (classes or schools) with unusual performance in a test. Their approach involves the following steps:

- for the test on which groups with unusual performance are to be investigated, each individual is classified into one of the J possible performance categories (eg based on their raw or scaled score and the performance cut scores)
- the predictor variables $(X_1, X_2, ..., X_I)$ are continuous (for example, the prior attainment or scores on different tests). Clark et al. initially proposed the approach for predicting current year's performance from previous year's test scores. However, the approach could be applied to situations where a suitable predictor or predictors are available)
- assuming a person with probabilities being classified into different performance categories $(Y_1, Y_2, ..., Y_J)$ to be $(\pi_1, \pi_2, ..., \pi_J)$, the cumulative probability that the person is classified into performance category *j* is modelled using the cumulative logit regression model:

$$\log\frac{P(Y \le j)}{1-P(Y \le j)} = \log\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_J} = \beta_{0j} + \sum_i \beta_{ji} x_i \qquad (75)$$

  where $\{\beta_{0j}\}$ and $\{\beta_{ji}\}$ are model parameters.

- for each person, the probabilities of being classified into individual performance categories are treated as expected values. The expected count of persons from group *k* that are classified into category *j*, $E(P_{jk})$, can be calculated as the sum of the probabilities of the persons in the class divided by the number of persons N_j in the group:

$$E(P_{jk}) = \frac{\sum_{n=1}^{N_k} \pi_{nj}}{N_k}$$

- for each performance category *j* for group k, work out the standardised residual as the difference between the observed proportion $P_{jk,0}$ and the expected proportion $E(P_{jk})$ divided by the standard error (partly related to the number of persons being classified into the category N_{jk}):

$$R_{jk} = \frac{P_{jk,0} - E(P_{jk})}{\sqrt{E(P_{jk})[1-E(P_{jk})]/N_{jk}}} \qquad (76)$$

The standardised residuals are expected to be normally distributed with a mean of 0 and a standard deviation of 1. At a specific performance level for the group, positive

values indicate higher than expected proportion of students from the group that were classified into the category, negative residual lower than expected proportion of students being classified into the category. Groups with values of the standardised residuals greater than 3 could be treated as performing significantly differently from expected.

Clark et al. (2013, 2017) suggested that particular attention may need to be paid to groups with extremely large positive residuals. Using simulated data, they showed that the cumulative logit regression model was considerably more effective than the weighted least squares regression method in identifying groups with unusual performance. The detection rate was found to be over 98% for the conditions simulated. The Type I errors were generally below the nominal levels.

The Regression and cluster based approaches

In many situations, a dependent variable on which outlying members (for example, classes or schools performed unusually on an achievement test) are to be identified is modelled using one or more independent variables. A member is flagged out as an outlier if its observed value on the dependent variable is significantly different from that predicted by the model. Simon (2014) suggested that many existing methods identify outlying schools with respect to all the schools included in the analysis and refers them as global outliers. He argues that schools with suspicious behaviour may not exhibit sufficient extremity to be identified as global outliers. But such schools may be regarded as outliers when compared with their peers – schools which are similar in many relevant aspects or similar values on the independent variables. He suggested that conventional techniques lack the ability to identify local outliers. Using data mining techniques, Simon developed an approach, the Regression based Local Outlier Detection algorithm (RedLOD), which can be used to identify groups which are local outliers with respect to a variable of interest. The basic assumption of this approach is that schools which have similar values on a set of independent variable should also be expected to have similar values on the dependent variable. This approach involves the following stages (see Simon, 2014):

- data preparation. For both dependent and independent variables, the raw data may be transformed onto different metrics and values for individual students are aggregated to produce school level data. It is the school level data that is used in the analysis.
- selection of independent variables. For the dependent variable, a set of intendent variables are selected. One of the approaches that can be used to select independent variable is through multiple regression. Contributions from the independent variables to the amount of variance in the independent variable that can be explained by the regression model can be examined.
- assessing the importance of independent variables and identifying peer schools. Once a set of independent variables have been selected, their relative importance

in identifying peers is determined through their weights which will be used to identify peer schools using the following steps:

1. Initialize all weights for the independent variables to w_k (k=1,2,…,K, where K is the total number of independent variables)
2. The weighted Euclidean distance between school i and j, D_{ij} , is calculated as:

$$D_{ij} = \sum_{k=1}^{K} w_k (x_{ki} - x_{kj})^2 \qquad (77)$$

   where x_{ki} and x_{kj} are the values of schools i and j on independent variable x_k
3. For each school s, form its current peer group P_g by selecting the closest pre-determined number of schools. Perform regression analysis using the data from the schools in the peer group and obtain the coefficients. Normalize the coefficients so that the absolute values of the coefficients sum to 1. The coefficient of the k-th independent variable for school is denoted as C_{sk} .For each school, a set of regression coefficients are obtained.
4. The weight w_k for the k-th independent variable is then recalculated as the mean of the corresponding coefficients of all schools (total number of S_g):

$$w_k = \frac{1}{S_g} \sum_{s=1}^{S_g} C_{sk}$$

   The weight for an independent variable is therefore related to its ability in predicting the dependent variable within peer groups.
5. Repeat steps 2-4 until the sum of the squares of the difference between the coefficients for two consecutive iterations is less than a pre-specified threshold. A value of 0.001 was used by Simon (2014).

Practical implementation of the above procedure may need to consider computational implications. In his study, Simon used 100 randomly selected schools to estimate weights. A value of 0.03 was use as the distance to identify peer schools.

- identifying local outliers within the peer groups. For the dependent variable on which outlying schools are to be identified, empirical p- value derived using bootstrap resampling with replacement can be used. The following steps will need to be taken:
  - for each school s, draw a bootstrap empirical distribution of the dependent variable from its peer schools (the school s itself is excluded). For each bootstrap sample, a p-value is calculated for the school. Repeat the sampling and work out the average p-value over the bootstrap samples. A small p-value would indicate that the school performed unusually better than its peers.

- flag schools with small p-values (eg p<=0.05) or schools with a low number of peer schools (less than 10)

Simon used the RegLCD approach to investigate local outliers for large scale real tests in a number of subjects and across several grades. He compared his method with other methods used to identify global outliers and showed that it was able to identify outlying schools which were missed by the other methods.

Multilevel modelling

The relationship between the test being investigated and the variables concerned can be modelled using linear regression models, including multilevel regression models. Multilevel models have been used in value added analysis extensively and also used for statistical moderation of results from school-based assessments in a number of countries (See, Kim and Lalancette, 2013; Hong Kong Examinations and Assessment Agency, 2012). He and Stockford (2015) proposed to use a two-level linear regression model with random intercept and fixed slope effects to identify schools (or test centres) that might have performed unusually on school-based non-exam assessment (NEA) components in relation to their performance on external exam (EE).

This model can be expressed as:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \end{aligned} \qquad (78)$$

where:

Y_{ij} : the score of student i from centre j on the school-based NEA

X_{ij} : the score of student i from centre j on the external exam assessment

γ_{00} : model parameter representing the fixed effect component of the intercept

u_{0j} : random intercept component at centre level

β_{1j} : model parameter representing fixed effect for the slope of the regression line at centre level

ε_{ij} : student level random error or residual

Graphically, this model produces centre regression lines which are parallel (having the same slope β_{1j}) and intersect with the y-axis at different locations (or with different values for the random component u_{0j} of the intercept). Since only the intercept can take different values for different centres, differences in the intercept values would represent any systematic differences in scores awarded on the NEA

component between schools, taking into consideration the effect of exam assessment scores. Therefore, the intercept could be used to identify schools that might have performed unusually on the non-exam assessment component. Furthermore, since values of u_{0j} are centred on γ_{00} (the intercept of the average line that crosses students from all schools), it represents the departure of the intercept for the centre from the average of all centres.

The following steps would be needed if using the approach described above to identify centres for which further moderation process may need to be taken:

- analyse the data using the model specified above with a multilevel modelling software
- rank order schools based on their values of the random intercept component u_{0j}
- centres with values of u_{0j} greater than $3\times\sigma_{\beta_{0j}}$ (the standard error of the intercept estimate arising from the model fitting process) may be regarded as outliers

Given the complex nature of multilevel modelling, implementation of the approach described above for operational use will likely involve the use of specialised software packages or, at least, bespoke routines implemented in standard statistical software packages.

Simple linear regression analysis

While a multilevel level model may describe the relationship between variables with a hierarchical structure more accurately than conventional linear regression models, its implementation for operational use may be complicated. He and Stockford (2015) described a procedure involving the use of the simple linear regression approach to identify schools with unusual performance on the NEAs. The simple linear regression model, which is similar to the linear regression model proposed by Li et al. (2014), can be used to describe the relationship between the NEA component and the EE assessment component can be expressed as:

$$y_i = \beta_0 + \beta x + \varepsilon_i \qquad (79)$$

where:

y_i : the score of student i on the non-exam assessment

x_i : the score of student i on the exam assessment

ε_i : residual or random error

β and β_0 : model parameters representing the slope and intercept of the regression line respectively

Equation (79) can be applied to students from individual centres or students from all centres. Model parameters in conventional regression analysis can be estimated using the least squares method (minimising the sum of the squares of the residuals). When it is applied to students from all centres, the hierarchical structure of the data is ignored. When it is applied to individual centres, different values for the model parameters may be produced for different centres. Use of the simple linear regression model represented by Equation (79) for statistical moderation would need to assume that the relationship between the two variables are similar for all schools. That is, values of the model parameters should be the same across the schools, taking into consideration any statistical uncertainties associated with their estimation.

To use the simple linear regression analysis approach to identify centres with unusual performance on the non-exam assessment, Equation (79) can firstly be applied to students from all centres and the global model parameters are estimated, which are denoted as β_{All} and $\beta_{0,All}$ respectively. For each school, the relationship between the two variables is assumed to be linear and the slope of the line takes the same value as that of β_{All}, but the intercept can take different values for different schools (ie the relationships between the two variables for the schools are characterized by parallel lines, as in the case of the multilevel modelling approach discussed earlier). Therefore, for a specific centre *j*:

$$y_{ji} = \beta_{All} + \beta_{j0} x_{ji} + \varepsilon_{ji} \qquad (80)$$

And the intercept β_{j0} for centre *j* can be calculated from $\beta_{j0} = \bar{y}_j - \beta_{All}\bar{x}_j$ where $\bar{y}_j$ is the average score of students from centre *j* on the non-exam assessment, and $\bar{x}_j$ is the average score of the students on the exam assessment. This ensures that the line crosses the centre of the data points from the centre. The standard error $\sigma_{\beta_0 j}$ of the intercept estimate β_{j0} may be estimated using the following equation:

$$\sigma_{\beta_0 j} = \sigma_j \sqrt{\frac{1}{N_j} + \frac{N_j \bar{x}_j^2}{\sigma_{j,x}^2}}$$

where N_j, σ_j and $\sigma_{j,x}$ are the number of students in centre *j*, the standard error of estimate (the standard deviation of the residuals) and the standard deviation of the predictor respectively. σ_j and $\sigma_{j,x}$ are defined as:

$$\sigma_j = \frac{\sum_i [y_{ji} - (\beta_{j0} + \beta_{All} x_{ji})]^2}{N_j - 2}$$

$$\sigma_{j,x}^2 = \frac{\sum_i (x_{ji} - \bar{x}_j)^2}{N_j}$$

Because the lines have the same slope, any systematic difference in the relationship between the two variables among the centres will be reflected only by the difference in the values of the intercept parameter β_{j0}. Centres with high intercept values performed better on the NEA than centres with lower values although their performance on the exam assessment may be similar. To make comparison between centres more meaningful and easier, a mean value of the centre intercept can be calculated and the centre value can then be compared with this mean value.

To use the procedure described above to identify schools that may have performed unusually on the NEA, the following steps would need to be taken:

- model the relationship between the two variables with Equation (79) using students from ALL centres to estimate the global model parameters (β_{All} and $\beta_{0,All}$)
- for each centre, assume that a linear relationship between the two variables exists and the slope of the line is the same as the global slope β_{All}
- work out the intercept of the line for the centre β_{j0}
- calculate the mean of the centre intercept values which can be denoted by $\bar{\beta}_{0Centre}$ (or use the global intercept value $\beta_{0,All}$ obtained using students from all centres)
- calculate the difference $\Delta\beta_{j0}$ between the centre intercept value and the mean for all centres (ie $\Delta\beta_{j0} = \beta_{j0} - \bar{\beta}_{0Centre}$). Centres with positive difference values performed better than the average performance of all centres on the NEA, taking into consideration their performance on the exam assessment. In contrast, centres with negative values performed below the average performance of all centres on the NEA. Centres with the absolute value of the difference $\Delta\beta_{j0}$ greater than three times the standard error ($\sigma_{\beta_0 j}$) of the centre intercept estimate may be regarded as outliers

The procedure described mirrors the multilevel modelling approach closely but represents a simplified version that would, arguably, be easier to implement for operational use. No specific software packages would have to be used for such an implementation, however, it is still necessary to fit a statistical model. In summary, this trades some statistical model fitting complexity for additional steps in the processing and a reduction in statistical power.

A residual analysis approach using standardised scores and principal axis

If Equation (79) is applied to students from all centres, a residual analysis or "value added" approach could also be developed and used for identifying centres which may be regarded as outliers (see He and Tymms, 2014; He and Stockford, 2015). The "value added" relationship considered here is between students' EE performance and their NEA mark. (Note that the term "value added" as being used here differs from its frequent use in the field relating prior attainment to grade outcome). These would be centres with exceptionally high (or low) values on the value added measure. To make a comparison between centres, instead of using Equation (79) to represent the relationship between the performance on the NEA and that on the EE, the principal axis may be used. For a bivariate dataset, such as that considered here, the principal axis is the line of symmetry on which the variance of the data points is maximised. When measures on both variables are standardised to have the same mean and standard deviation, the principal axis is reduced to the identity line:

$$Y_i = X_i + R_i \tag{81}$$

where Y_i and X_i are the standardized scores of student *i* on the non-exam and exam assessments respectively, and R_i is the residual. Equation (81) suggests that, for student *i*, given his/her observed score X_i on the exam assessment, his/her expected or predicted score on the NEA (the average score on the NEA for students with similar exam scores) would be also X_i. When the residual R_i $(=Y_i - X_i)$ is positive, the student performed better on the non-exam assessment than the average performance of students with similar exam performance. If the residual is negative, the student performed below the average of the students with similar level of exam performance.

For centre *j* with N_j students, the average value added $\overline{VA}_j$ is calculated as:

$$\overline{VA}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} R_i = \frac{1}{N_j}\sum_{i=1}^{N_j}(Y_i - X_i) \tag{82}$$

This average value added may be used as a measure for quantifying any systematic difference in performance on the NEA between centres, taking into consideration the performance on the exam assessment. As the average value added for all centres is zero, centres with positive value added performed better on the NEA than the average performance of centres with similar exam performance. Centres with exceptionally high positive or negative "value added" measures could be regarded as outliers.

This approach would involve the following steps:

- standardize both sets of raw scores (NEA and EE) to have the same mean and standard deviation, for example, 0 and 1.0 respectively.
- for each student, work out their value added score $R_i = Y_i - X_i$ (based on the standardised scores) to create a set of value added scores
- for each centre, work out their mean value added score $\overline{VA}_j$
- work out the standard deviation of the centre level value added scores σ_{VA}
- if the absolute value of the centre level value added $\overline{VA}_j$ is greater than $3 \times \sigma_{VA}$, then the school may be treated as outliers

This procedure is easy to implement for operational application as it does not require the fitting of any statistical models and would not need use of any specialised or statistical software packages.

5. Concluding remarks

A range of statistical techniques have been developed to study anomaly in results from high-stakes tests and assessments. These generally involve statistical test of significance in difference between the observed item response patterns or test scores from test-takers and those expected from theoretical/empirical models or the responses and scores from other test-takers in the sample or population. Such significance test involves the analysis of:

- response patterns on items from individual test-takers in relation to those expected from theoretical/empirical models or from other test-takers
- item responses and test score distributions within groups of test-takers in relation to those expected or from other groups
- relationship between performances on different subsets of items in the test for individuals and groups
- relationship between the performance on the test and performance on other variables for individuals and groups

Many of the methods reviewed have been developed or used to detect anonymous responses and scores from high-stakes tests and examinations associated with cheating or test collusion at individual test-taker level or group level. Cheating in high-stakes testing can take various forms, including test-taker cheating; teacher cheating, test coaching, either by a classroom teacher or from a review course; systematic answer sharing during the test; use of harvested items; inappropriate marking or scoring of test-takers' work; and others (see Wollack and Mayes, 2011; Belov, 2013). Cheating and collusion have been a concern for assessment providers, the relevant authorities, and other users of test results. With the rapid advance in technology, new techniques are being increasingly used in cheating, which makes cheating more sophisticated and difficult to detect using conventional means, and statistical approaches can provide useful information. It should however

be emphasised that there can be many other factors other than cheating that can produce anomalous responses and test scores. For example, if the test is inappropriate for the test-takers being tested in terms of the levels of ability and the type of skills and knowledge being assessed by the test, or the test-taker behaves unconventionally when answering questions (eg random guessing, language deficiency, creative interpretation of test items), anomalous responses and scores may result (see Meijer, 1996a, b; Karabatson, 2003; Thiessen, 2008). As Bishop and Stephens (2013) suggested, statistical techniques used to detect cheating behaviours can identify statistically unusual patterns in test data. Although they can provide some kind of likelihood-based conclusion about possible cheating for those who are interested in the performance of the test and use of the results, it is impossible for them to prove that cheating or test collusion has actually happened. Rather, they demonstrate how extremely unlikely the identified anomalous results would happen based on the given underlying assumptions made about the models used to analyse the test data.

It is also worth noting that the same test data may be analysed using different methods to detect the same or different aberrant behaviours. In situations where different methods can be used to the whole dataset or parts of the dataset, application of multiple methods may be beneficial as different methods examine the data from different perspectives. A high percentage of respondents simultaneously flagged by several aberrant indices could be an indication of aberrant response behaviour (eg Meijer and Tendeiro, 2014; Plackner and Primoli, 2014). Further, multiple methods may also be used to identify the extent to which different methods account for variation in detecting test-taking irregularities associated with test collusion (see Plackner and Primoli, 2014).

References

Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires, *British Journal of Mathematical and Statistical Psychology* 31, 84-98.

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association* 69, 44–49.

Angrist, J. D., Battistin, E. and Vuri, D. (2014). In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno. NBER Working Paper 20173.

Armstrong R. D., Stoumbos, Z. G., Kung, M. T. and Shi, M. (2007). On the Performance of the l_z person-fit Statistic *Practical Assessment Research & Evaluation* 12. Available online: http://pareonline.net/getvn.asp?v=12&n=16

Armstrong, R. and Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement* 46, 408-428.

Battistin, E., De Nadai, M. and Vuri, D. (2014). Counting rotten apples: Student achievement and score manipulation in Italian elementary schools. The Institute for the Study of Labor (IZA). Available online at: http://ftp.iza.org/dp8405.pdf

Belov, D. (2015). Robust detection of examinees with aberrant answer changes. *Journal of Educational Measurement* 52, 437–456.

Belov, D. (2016). Comparing the performance of eight Item preknowledge detection statistics. *Applied Psychological Measurement* 40, 83-97.

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement* 35, 495–517.

Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement* 50, 141–163.

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing* 2, 37–58.

Belov, D. I. and Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement* 34, 379–392.

Belov, D. I., Pashley, P. J., Lewis, C. and Armstrong, R. D. (2007). Detecting aberrant responses with Kullback–Leibler distance. In K. Shigemasu, A. Okada,

T. Imaizumi, & T. Hoshino (Eds.), New trends in psychometrics (pp. 7-14). Universal Academy Press: Tokyo, Japan.

Bezdek J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press: New York.

Bishop, K. L. and Stephens, C.N. (2013). Detecting Unusual Item Response Patterns Based on Likelihood of Answer Paper presented at the Statistical Detection of Potential Test Fraud Conference. Madison, Wisconsin.

Bliss, T. (2012). Statistical methods to detect cheating on tests: a review of the literature. Brigham Young University.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 46, 443–459.

Bradlow, E. T., Weiss, R. E. and Cho, M. (1998). Bayesian identification of outliers in 1 computerized adaptive testing. *Journal of the American Statistical Association* 93, 910-919.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Lawrence Erlbaum Associates: Mahwah, NJ.

Cizek, G. and Wollack, J. (2017). Exploring cheating on tests – the context, the concern, and the challenges. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 3-19. Routledge: New York.

Clark, J. M. (2012). Nested factor analytic model comparison as a means to detect aberrant response patterns. Paper presented at the Statistical Detection of Potential Test Fraud Conference. Lawrence, KS.

Clark, J. M., Skorupski, W. P. and Murphy, S. T. (2013). Using non-linear regression to identify unusual performance level classification rates. Paper presented at the Statistical Detection of Potential Test Fraud Conference. Madison, Wisconsin.

Clark, J. Skorupski, W. and Murphy, S. (2017). Using nonlinear regression to identify unusual performance level classification rates. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 245-261. Routledge: New York.

Clark, J. M., Skorupski, W., Jirka, S., McBride, M., Wang, C. and Murphy, S. (2014). An investigation into statistical methods to identify aberrant response patterns. Research Report, Pearson. Available online at:

https://www.google.co.uk/#q=An+investigation+into+statistical+methods+to+identify+aberrant+response+patterns

Donlon, T. F. and Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement* 28, 105-113

Drasgow, F., Levine, M.V. and Williams, E.A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology* 38, 67-86.

Egberink, I., Meijer, R., Veldkamp, B., Schakel, L. and Smid, N. (2010). Detection of aberrant item score patterns in a computerized adaptive test: An empirical example using the CUSUM. *Personality and Individual Differences* 48, 921-925.

Emmen, P. (2011). A Person-Fit Analysis of Personality Data. Master Thesis. Department of Social and Organizational Psychology. Vrije Universiteit.

Emons, W. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement* 32, 224-247.

Fernández, S. P. (2016). A new method for the correction of test scores manipulation. Discussion Paper, Bank of Italy. Avaiable online at: https://www.bancaditalia.it/pubblicazioni/temi-discussione/2016/2016-1047/en_tema_1047.pdf?language_id=1

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research* 42, 481-507.

Ferrando, P. J. (2009). Multidimensional factor-analysis-based procedures for assessing scalability in personality measurement. *Structural Equation Modeling* 16, 109-133.

Frary, R. B., Tideman, T. N. and Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics* 6, 152–165.

Gaertner, M. and McBride, Y. (2017). Detecting unexpected change in pass rates – a comparison of two statistic approaches. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 262-279. Routledge: New York.

Guo, J. and Drasgow, F. (2010). Identifying Cheating on Unproctored Internet Tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment* 18, 351-364.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and Precision* (pp. 60–90). Princeton University Press: Princeton, NJ.

Hambleton, R., Swaminathan, H. and Rogers, H. (1991). *Fundamentals of item response theory*. Sage Publications: London, England.

Harnisch, D. L. and Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133–46.

He, Q. and Tymms, P. (2014). The principal axis approach to value added calculation. *Educational Research and Evaluation* 20, 25-43.

He, Q. and Stockford, I. (2015). Possible statistical techniques for identifying centres with unusual performance on non-exam assessment in GCSE Computer Science for moderation. Ofqual Internal Report.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Tech. Rep. No. 96–4). Educational Testing Service: Princeton, NJ.

Hong Kong Examinations and Assessment Agency (HKEAA) (2012). Moderation of School-based Assessment Scores in the HKDSE. Available online at: http://www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/HKDSE-SBA-ModerationBooklet_r.pdf

Huang, T. W. (2012). Aberrance detection powers of the BW and person-fit indices. *Educational Technology and Society* 15, 28–37.

Jacob, B. A. and Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. The Quarterly Journal of Economics, 118(3):843{877.

Kane, M. T. and Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement* 4, 105-126

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education* 16, 277-298.

Keeves, J. and Alagumalai, S. (1999). Item banking. In *Advances in measurement in educational research and assessment* ed. G. Masters and J. Keeves, 23-42. Elsevier Science: The Netherlands.

Kim, H. and Lalancette, D. (2013). Literature review on the value-added measurement in higher education. OECD. Available online at: http://www.oecd.org/edu/skills-beyond-school/Litterature%20Review%20VAM.pdf

Kolen, M. J. and Brennan, R. L. (2008). *Test equating, scaling, and linking: Methods and practices* (Second ed.). Springer-Verlag: New York.

Levine, M. V. and Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics* 4, 269–290.

Li, M. F. and Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement* 21, 215-231.

Li, X, Huang, C. and Harris, D. (2014) .Examining Individual and Cluster Test Irregularities in Mixed-Format Testing. Paper presented at the 2014 Conference on Test Security. Iowa City, Iowa.

Linacre, J. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions* 16, p.878.

Lord, F. M. and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ‘‘equatings’’ *Applied Psychological Measurement* 8, 452–461.

Madaus, G., M. Russell and J. Higgins (2009). *The Paradoxes of High Stakes Testing: How they affect students, their parents, teachers, principals, schools, and society*. Information Age Publishing: Charlotte, CN, USA.

Magis, D., Raîche, G. and Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics* 37, 57-81.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149-74.

Maynes, D. (2005). *M4 — A new answer-copying index*. Unpublished manuscript, Caveon Test Security, Midvale, UT.

Maynes, D. (2014a). Detection of non-independent test taking by similarity analysis. In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 52-80. Routledge Research in Education: New York

Maynes, D. (2014b). A method for measuring performance inconsistency by using score differences. In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 186-199. Routledge Research in Education: New York

McClintock, J. (2015). Erasure analyses: reducing the number of false positives. *Applied Measurement in Education* 28, 14-32,

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement* 18, 311–314.

Meijer, R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement 39,* 219-233.

Meijer, R. R. (Guest Ed.). (1996a). Person fit research: Theory and applications [Special Issue]. *Applied Measurement In Education, 9*(1).

Meijer, R. R. (1996b). Person-fit research: An introduction. *Applied Measurement in Education, 9,* 3–8.

Meijer, R. R. and Sijtsma, K. (2001). Methodology review: evaluating person fit. *Applied Psychological Measurement* 25, 107-135

Meijer, R. R. and Tendeiro, J. N. (2014). *The use of person-fit scores in high stakes educational testing: How to use them and what they tell us* (LSAC Research Report 14-03). Retrieved from http://www.lsac.org/lsacresources/research/all/rr

Molenaar, I. W. and Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika* 55*,* 75–106.

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* 16, 159-176.

Office for Standards in Education (Ofsted) (2012a). *Mathematics: Made to Measure*. Available online at: www.ofsted.gov.uk/resources/mathematics-made-measure.

Office for Standards in Education (Ofsted) (2012b). *Moving English Forward*. Available online at: www.ofsted.gov.uk/resources/moving-english-forward.

Office of Qualifications and Examinations Regulation (Ofqual) (2012). *GCSE English 2012*. Available online at: www.ofqual.gov.uk/files/2012-11-02-gcse-english-final-report-and-appendices.pdf

Plackner, C. and Primoli, V. (2014). A compare-and-contrast analysis of multiple methods. In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 203-219. Routledge Research in Education: New York.

Quintano, C., Castellano, R. and Longobardi, S. (2009). A Fuzzy Clustering Approach to Improve the Accuracy of Italian Student Data. An Experimental

Procedure to Correct the Impact of the Outliers on Assessment Test Scores. Statistica and Applicazioni, Vol.VII(2), 149–171.

Rasch, G. (1960). *Probabilitistic models for some intelligence and attainment tests*. Danish Paedagogiske Institute: Copenhagen, Denmark.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement* 19, 213-229.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* 17.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meishi Tosho.

Shu, Z., Henson, R. And Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika* 78, 481-497.

Sijtsma, K. and Meijer, R. (2001). The person response function as a tool in person-fit research. *Psychometrika* 66, 191-208.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitative Methoden* 7, 131-145.

Simon, M. (2014). Local outlier detection in data forensics: data mining approach to flag unusual schools. In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 83-100. Routledge Research in Education: New York.

Sinharay, S. (2015) Assessing person fit using l^*_z and the posterior predictive model checking method for dichotomous item response theory models. *International Journal of Quantitative Research in Education* 2, 265-

Sinharay, S. (2015) Assessment of Person Fit for Mixed-Format Tests. *Journal of Educational and Behavioral Statistics* 40, 343-365

Skorupski, W. and Egan, K. (2012). A hierarchical linear modeling approach for detecting cheating and aberrance. Paper presented at the May, 2012 Conference on the Statistical Detection of Potential Test Fraud. Lawrence, KS.

Skorupski, W. and Egan, K. (2014). A Bayesian hierarchical linear modelling approach for detecting cheating and aberrance. . In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 121-133. Routledge Research in Education: New York.

Skorupski, W., Fitzpatrick, J. and Egan, K. (2017). A Bayesian hierarchical model for detecting aberrant growth at group level. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 232-244. Routledge: New York.

Smith, R. and Davis-Becker, S. (2011). Detecting Suspect Examinees: An Application of Differential Person Functioning Analysis. Paper presented at the National Council on Measurement in Education Annual Conference. New Orleans, LA.

Snijders, T. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika* 66, 331-342.

Sotaridona, L. S. and Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying in a multiple-choice test. *Journal of Educational Measurement* 39, 115–132.

Sotaridona, L. S. and Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement* 40, 53–70.

Sotaridona, L. S., Wibowo, A. and Hendrawan, I. (2014). A parametric approach to detect a disproportionate number of identical item responses on a test. In Kingston, N. M. and Clark, A. K. (Editors): *Test Fraud: Statistical Detection and Methodology*, pp 38-52. Routledge Research in Education: New York.

Sung, H. and Kang, T. (2006). Choosing a polytomous IRT model using Bayesian model selection methods. Paper presented at National Council on Measurement in Education Annual Meeting. San Francisco, CA.

Tan, J. and Yates, S. (2007). A Rasch analysis of the Academic Self-Concept Questionnaire. *International Education Journal* 8, 470-484.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika* 49, 95–110.

Tatsuoka, K. K. and Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics* 7, 215-231.

Tatsuoka, K. K. and Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement* 20, 221-230.

Tendeiro, J. N. (2015). PerFit (version 1.3) [Computer software]. University of Groningen. Retrieved from http://r-forge.r-project.org/R/?group_id=1878

Tendeiro, J. N. and Meijer, R. R. (2014). Detection of invalid test Scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement* 51, 239-259.

Tendeiro, J. N., Meijer, R.R. and Niessen, A. S. (2016). PerFit: An R Package for person fit in IRT. *Journal of Statistical Software* 74, issue 5.

Thiessen, B. (2007). Case study—policies to address educator cheating. Avalaible online at: http://www.bradthiessen.com/html5/docs/format.pdf

Thiessen, B. (2008). Relationship between test security policies and test score manipulations. PhD Thesis, University of Iowa. Available online at: http://www.bradthiessen.com/html5/docs/thiessen.pdf

Trabin, T. E. and Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing*. Academic Press: New York.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology* 13, 267–298.

van der Linden, W. J. and Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics* 37, 180–199.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. InW. J. van der Linden & C. A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Kluwer Academic Publishers: Dordrecht, the Netherlands.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics* 26, 199–218.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous Items. *Applied Psychological Measurement* 26**,** 164-180.

Victorian Curriculum and Assessment Agency (VCAA) (2016). Statistical Moderation of VCE Coursework. Available online at: http://www.vcaa.vic.edu.au/Pages/vce/exams/statisticalmoderation/statmod.aspx

Wesolowsky G. (2000a) Detecting Excessive Similarity in Answers on Multiple Choice Exams. *Journal of Applied Statistics* 27, 909-921.

Wesolowsky, G. (2000b). *Statistical detection of cheating (copying, collusion) on multiple choice tests and examinations. Avaiable online at:* http://www.business.mcmaster.ca/msis/profs/wesolo/wesolo.htm

Wibowo, A., Sotaridona, L. and Hendrawan, I. (2013). *Statistical models for flagging unusual number of wrong-to-right erasures.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Wollack, J. A. and Maynes, D. (2011). Detection of test collusion using item response data. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Wollack, J. A., (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement* 21, 307-320.

Wollack, J. A., (2004). Detecting Answer Copying on High-Stakes Exams. *The Bar Examiner* 73, 35-45.

Wollack, J. and Maynes, D. (2017). Detection of test collusion using cluster analysis. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 25-46. Routledge: New York.

Wollack, J. A., (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education* 19, 265-288

Wong, H., McGrath, C. and King, K. (2011). Rasch validation of the early childhood oral health impact scale. *Community Dent Oral Epidemiology* 39, 449–457.

Wright, B. and Masters, G. (1982). *Rating scale analysis: Rasch measurement*. MESA Press: Chicago, USA.

Wright, B. and Stone, M. (1979). *Best test design: Rasch measurement*. MESA Press: Chicago, USA.

Wright, B.D. and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement* 29, 23-48.

Wu, M. and Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions: Melbourne.

Zhang, Y., Searcy, C. A. and Horn, L. (2011). Mapping clusters of aberrant patterns in item responses. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Zopluoglu, C. (2016a). Classification performance of answer-copying indices under different types of IRT models. *Applied Psychological Measurement* 40, 592–607.

Zopluoglu, C. (2016b). Package 'CopyDetect' - computing statistical indices to detect answer copying on multiple-choice tests. Available online at: https://cran.r-project.org/web/packages/CopyDetect/CopyDetect.pdf

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance – understanding the status quo. In Cizek, G. and Wollack, J. (Editors): *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pp 25-46. Routledge: New York.

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346