# ANALYZING MISSINGNESS AT THE LEVEL OF THE INDIVIDUAL RESPONDENT: COMPARISON OF FIVE STATISTICAL TESTS

PASQUALE ANSELMI
EGIDIO ROBUSTO
FRANCESCA CRISTANTE
UNIVERSITY OF PADOVA

To date, missingness has been investigated at the level of the entire sample of respondents, on data samples deriving from a unique mechanism (either MCAR, or MAR, or MNAR). In this work, missingness is investigated at the level of the individual respondent, on data samples consisting of patterns of missingness deriving from different mechanisms. Five tests (runs test, Wilcoxon rank sum test, point-biserial correlation, *t*-test, standardized outfit) are used for detecting whether the particular pattern of missingness displayed by an individual respondent has been generated by a random (MCAR, MAR) or nonrandom (MNAR) mechanism. For all the tests, the rejection of the null hypothesis for a particular pattern is taken as an indication that the pattern might have been caused by a nonrandom mechanism. A simulation study and a real data application example showed that Wilcoxon rank sum test and *t*-test outperform the other tests in identifying the MNAR patterns.

Key words: Missing data; MCAR; MAR; MNAR; Response behavior.

*Correspondence concerning this article should be addressed to Pasquale Anselmi, Department FISPPA – Section of Applied Psychology, University of Padova, Via Venezia 14, 35131 Padova (PD), Italy. Email: pasquale.anselmi@unipd.it*

Missing data are a common problem in a variety of measurement settings, including responses to items on both cognitive and affective assessments. There are many reasons underlying missing data, often unknown to researchers. The respondent may have missed one or more items either inadvertently or because he/she did not know the answer and was afraid to guess. Again, the respondent may have become bored while filling in the questionnaire or the test, and left some items unanswered. Moreover, he/she may have felt some items to be embarrassing, threatening, or intrusive to privacy, thus preferring to skip them.

Following Rubin (1976), missing data are said to be *missing at random* (MAR) if the cause of missingness does not depend on the missing values themselves, but depends on another variable included in the analysis. If missingness is also unrelated to any variable in the analysis, the missing data are defined *missing completely at random* (MCAR). Missing data are *missing not at random* (MNAR) if nonresponse depends on the missing values themselves. When missingness is MCAR, it is ignorable because the observed data are a random sample from the complete data sample (Little & Rubin, 1987). Missingness is also ignorable when it is of the MAR type because, in this case, the observed data are a random sample from subsamples of the complete data. For instance, most women may refuse to answer an item that is offensive or sensitive to women but that does not affect men in the same way. Thus, the distribution of item scores will be different between women and men, but it will be the same for respondents and nonrespondents in both groups. Missingness of the MNAR type is nonignorable because the observed data are not a random sample from the complete data sample or from subsamples.

When missing data occur for reasons beyond our control, we must make assumptions about the mechanism that caused them. This is important because the adoption of less than optimum strategies for handling the missing values can lead to biased estimates, distorted statistical power, and invalid conclusions (Acock, 2005; Osborne, 2013). MCAR or MAR data can be problematic from a power perspective (in that sample size decreases and standard errors increase), but they would not potentially bias the results (the available data are representative of the population). Conversely, MNAR data could potentially be a strong biasing influence (Rubin, 1976; Sijtsma & van der Ark, 2003; Stuart, Azur, Frangakis, & Leaf, 2009). For instance, if some items are more often missed by worst-performing students than by best-performing students, there would be an overestimation of average performance, due to more missing data from lower-performing students, and an underestimation of the standard deviation, due to less dispersion at the lower extreme of the distribution (Osborne, 2013).

The inappropriate treatment of missing data could lead researchers to draw invalid conclusions about the nature of relationships existing in the population. Stuart et al. (2009) provided interesting examples in the context of a national health survey. In one example, eliminating cases with missing data led to the conclusion that individuals who start smoking earlier in life are more emotionally strong and less functionally impaired than individuals who start smoking later — a finding that is contrary to research. In another example, individuals who drink more had fewer internalizing problems (e.g., depression, anxiety) — another finding that is not in line with research. After appropriate handling of missing data, these relationships were more consistent with the literature. Osborne (2013) analyzed data from a large educational longitudinal study and showed that the strong positive correlation between math achievement scores and reading scores ($r$ = .77) became weak negative ($r$ = –.20) after a MNAR decimation mechanism was applied to the original data sample.

In practical situations, the researcher has to decide whether the missingness at hand is ignorable or not on the basis of the pattern of missingness in the data. Huisman (1999) proposed a statistical test for investigating whether missingness in a data sample is random or not. A Pearson's Chi-square statistic is computed, which compares the frequency counts of the observed missing scores with that expected under the assumption of random missingness. Other tests provide information about missingness at the item level (see, e.g., Osborne, 2013). A missingness variable pertaining to a certain item (0 = missed, 1 = responded) can be correlated with substantive or ancillary variables (e.g., socioeconomic status, test scores). A *t*-test can be performed to investigate if individuals with missing data on one variable are significantly different on other similar variables. It is worth noting that, to date, the analysis of missingness has been carried out at the level of the entire sample of respondents, on data samples deriving from a unique missing mechanism (MCAR, MAR, or MNAR).

The present work represents a sharp departure from the traditional way missingness is investigated. A first novelty is that missingness is studied at the level of the individual respondent, instead of being considered at the level of the entire sample of respondents. The particular pattern of missingness displayed by a respondent is analyzed in order to infer if it has been caused by a random or nonrandom mechanism. Such an analysis is useful whenever we want to make proper decisions at the level of the individual respondent. For instance, examinees might skip the items that they are not able to respond. If this is the case, recoding the missing responses as incorrect would be appropriate (Weeks, von Davier, & Yamamoto, 2016). However, omitting does not only depend on examinee's ability, but it also involves a "dimension of temperament" (Lord, 1983). For instance, in particular conditions, examinees with low self-esteem or self-efficacy might omit responses to items that otherwise they may have been able to respond (Matters & Burnett, 2003). If this is the case, recoding the missing responses as incorrect responses would introduce a negative bias. The biases resulting from an inappropriate recoding of missing responses might cause untrustworthy test results and detrimental effects on the choices that are based on them (e.g., student admission, personnel selection). After determining that a particular pattern of missingness was caused by a random or nonrandom mechanism, the researcher can select the appropriate analysis method.

A second novelty of the present work is the analysis of mixed data samples, each one consisting of patterns of missingness deriving from different mechanisms (i.e., patterns deriving from a MCAR mechanism, patterns deriving from a MAR mechanism, and patterns deriving from a MNAR mechanism). There can be situations in which it is reasonable to assume that most of the patterns of missingness in the data sample derive from the same mechanism, as well as situations in which such an assumption can not be made. An example of the former case is high-stakes achievement tests, in which missing data might be mostly due to low-ability respondents who skip the items that they are not capable of solving. Thus, the patterns of missingness are mostly caused by a nonrandom mechanism. An example of the latter case is large-scale educational surveys (e.g., PIRLS, PISA, TIMSS), in which students do not receive individual scores and the outcomes of the assessment are inconsequential for them. The low-stakes nature of these surveys, together with individual differences in the way of completing the task (achievement motivation, perceived self-efficacy, commitment; see, e.g., Matters & Burnett, 2003), might contribute to the observance of patterns of missingness that derive from a nonrandom mechanism and patterns of missingness that derive from a random mechanism. It might be very difficult to make assumptions about the relative frequency of these two types of patterns of missingness in the overall data sample.

The article presents five tests for analyzing the specific pattern of missingness displayed by an individual respondent: runs test, Wilcoxon rank sum test, point-biserial correlation, *t*-test, and standardized outfit. These tests are common in classical testing theory and item response theory and, to our knowledge, they have never been used to analyze missingness at the level of the individual respondent. The five tests are computed on the pattern of missingness displayed by the respondent. For all the tests, the null hypothesis specifies that the mechanism underlying the pattern of missingness is random. The rejection of the null hypothesis suggests that the pattern might have been caused by a nonrandom mechanism.

## Tests for Analyzing the Pattern of Missingness Displayed by an Individual Respondent

The five tests presented in this section are: runs test, Wilcoxon rank sum test, point-biserial correlation, *t*-test, and standardized outfit. Some notation is introduced here, that will be used to describe the tests. Let “R” denote the items that were responded by the individual, and let “M” denote the items that were missed by him/her. Let $N_R$ and $N_M$ be respectively the number of items that were responded and missed by the individual, and let $N = N_R + N_M$ be the total number of items. A terminology will be used that refers to cognitive assessments (e.g., easy and difficult items, correct and incorrect responses). However, the presented tests can be applied also in other assessment contexts.

### The Runs Test

The runs test (Bradley, 1968; Wald & Wolfowitz, 1940), also called Wald-Wolfowitz test, is a nonparametric statistical test that checks the null hypothesis that a sequence of two-valued elements is random. It is based on the computation of the number of runs in the sequence, each of which is a segment of the sequence consisting of consecutive equal elements. For instance, the 20-element long sequence “AAAABBBAAABBAAAAAAAA” is characterized by 5 runs, 3 consisting of “A” and 2 consisting of “B.”

In the present study, runs are computed on the sequence of the items that were responded "R" or missed "M" by an individual respondent, after the items have been ordered by increasing difficulty. Item difficulty is represented by the proportion of incorrect responses to the item over the number of responses given to that item. Let us consider the following two sequences on a 20-item long test, both characterized by 9 missing responses:

RRMRRRMMRRMMRMRMMRRM
RRRRRRRRMRMMRRMMMMMM

$\longleftrightarrow$

Less difficult item More difficult item

The first sequence (12 runs) is assumed to be more probable than the second sequence (6 runs) under the null hypothesis that it was produced in a random manner.

Under the null hypothesis, the number of runs in a sequence of $N$ items is a random variable whose conditional distribution, given the observation of $N_\mathrm{R}$ responded items and $N_\mathrm{M}$ missed items, is approximately normal, with

$$\mu_r = \frac{2\,N_\mathrm{R}\;N_\mathrm{M}}{N} + 1, \tag{1}$$

and

$$\sigma_r^2 = \frac{2\,N_\mathrm{R}\,N_\mathrm{M}\,(2\,N_\mathrm{R}\,N_\mathrm{M} - N)}{N^2(N-1)}. \tag{2}$$

These parameters do not assume that responded and missed items have the same probability of occurring, but only assume that they are independent and identically distributed. If the number of runs is significantly lower than expected, the null hypothesis that the sequence is produced in a random manner may be rejected. In the example at hand, $\mu_r$ = 10.9 and $\sigma_r^2$ = 4.64 for both sequences. Considering a Type-I error probability of .05, the null hypothesis is retained for the first sequence $z$ = (12 – 10.9) / $\sqrt{4.64}$ = 51, $p$ = .70) and rejected for the second sequence ($z$ = −2.27; $p$ < .05).

## The Wilcoxon Rank Sum Test

The Wilcoxon rank sum test (Gibbons, 1985; Hollander & Wolfe, 1973) is a nonparametric statistical test of the null hypothesis that two samples of observations derive from populations with equal values. The test is based upon ranking all the observations from the two samples, without regard to which sample they belong to. The statistic $W$ is the sum of the ranks for the observations from one of the two samples.

In the present study, the Wilcoxon rank sum test is computed on the sequence of responded and missed items, after all the items have been sorted by increasing difficulty. Let us consider the following sequence of responded "R" and missed "M" items on a 20-item long test:

Less difficult item More difficult item

$\longleftrightarrow$

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item: | R | R | R | R | R | R | R | M | R | R | M | M | R | M | R | M | M | R | R | M |
| Ranks: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |

The example at hand represents a situation in which there are not items with the same difficulty level. When this is not the case, the same average rank is assigned to all the items with the same difficulty. The sum of the ranks for the missing responses is $W_{\mathrm{M}} = 98$.

Under the null hypothesis, the statistic $W_{\mathrm{M}}$ is approximately normally distributed, with

$$\mu_{W_{\mathrm{M}}} = \frac{N_{\mathrm{M}}\ (N+1)}{2}, \tag{3}$$

and

$$\sigma^2_W = \frac{N_{\mathrm{R}}\ N_{\mathrm{M}}\ (N+1)}{12}. \tag{4}$$

If the statistic $W_{\mathrm{M}}$ is significantly larger than expected, the null hypothesis that the responded and missed items do not differ in difficulty may be rejected. In the example at hand, $\mu_{W_{\mathrm{M}}} = 73.50$ and $\sigma^2_W = 159.25$. Considering a Type-I error probability of .05, the null hypothesis is rejected ($z = (98 - 73.50)/\sqrt{159.25} = 1.94$; $p < .05$).

## Point-Biserial Correlation

The point-biserial correlation coefficient (see, e.g., Glass & Hopkins, 1996) is a statistic used when one variable is continuous and the other variable is dichotomous. In the present study, the continuous variable is represented by the number of missing responses to each item across the entire sample. The point-biserial correlation coefficient $r_{\mathrm{pb}}$ is computed as

$$r_{\mathrm{pb}} = \frac{\bar{M}_{\mathrm{M}} - \bar{M}_{\mathrm{R}}}{s}\sqrt{\frac{N_{\mathrm{M}} - N_{\mathrm{R}}}{N^2}}, \tag{5}$$

where $\bar{M}_{\mathrm{M}}$ and $\bar{M}_{\mathrm{R}}$ are respectively the average number of missing responses to the items that were missed and responded by the individual, and $s$ is the sample standard deviation of the number of missing responses to all the items. The statistic $r_{\mathrm{pb}}$ expresses the strength of the association between the response behavior of the individual (i.e., response or omission to an item) and that exhibited by the overall sample.

A formal test of the null hypothesis that the correlation is 0 in the population can be accomplished by computing the statistic $t_{\mathrm{pb}} = r_{\mathrm{pb}}\sqrt{(N-2)/(1-r^2_{\mathrm{pb}})}$, which has an approximate Student's $t$ distribution with degrees of freedom equal to $N - 2$. The rejection of the null hypothesis suggests that the response behavior of the individual is consistent with that exhibited by the entire sample.

## *T*-Test

It is a parametric statistical test of the null hypothesis that two samples of observations derive from populations with equal means. In the present study, the *t*-test is used to compare the average difficulties of the items that were responded or missed by each individual. The rejection of the null hypothesis suggests that responded and missed items differ in difficulty.

If the responded and missed items derive from populations with unequal variance, the statistic *t* is computed as

$$t = \frac{\bar{D}_{\mathrm{M}} - \bar{D}_{\mathrm{R}}}{\sqrt{\frac{s_{\mathrm{M}}^2}{N_{\mathrm{M}}} + \frac{s_{\mathrm{R}}^2}{N_{\mathrm{R}}}}}, \tag{6}$$

where $\bar{D}_{\mathrm{M}}$ and $\bar{D}_{\mathrm{R}}$ are respectively the average difficulties of the missed and responded items, and $s_{\mathrm{M}}$ and $s_{\mathrm{R}}$ are their sample standard deviations. Under the null hypothesis, the statistic *t* has an approximate Student's *t* distribution with a number of degrees of freedom given by Welch-Satterthwaite's approximation (Satterthwaite, 1946; Welch, 1947). If the responded and missed items derive from populations with equal variance, the standard deviations $s_{\mathrm{M}}$ and $s_{\mathrm{R}}$ in the previous equation are replaced by the pooled standard deviation

$$s_{\mathrm{pooled}} = \sqrt{\frac{(N_{\mathrm{M}} - 1)s_{\mathrm{M}}^2 + (N_{\mathrm{R}} - 1)s_{\mathrm{R}}^2}{N - 2}}, \tag{7}$$

and the statistic *t* has Student's *t* distribution with $N - 2$ degrees of freedom.

## Standardized Outfit

The outfit mean-square (Wright & Masters, 1982) is a data-model fit statistic widely used in item response theory, which is an important theoretical framework for cognitive and personality assessment (see, e.g., Anselmi, Vidotto, Bettinardi, & Bertolotti, 2015; Colledani, Robusto, & Anselmi, 2018; Da Dalt et al., 2013, 2015; Thomas, 2011; Zanon, Hutz, Yoo, & Hambleton, 2016). Based on the comparison between observed and expected responses, it has the form of a $\chi^2$ statistic divided by its degrees of freedom. The expected value is 1. Values smaller than 1 indicate that the observations are too predictable (overfit), whereas values greater than 1 indicate unpredictability (underfit; Wright & Linacre, 1994).

In the present study, the outfit statistic is used to compare the actual response behavior of an individual respondent (i.e., response or omission to each item) with that expected according to the simple logistic model (SLM; Rasch, 1960). To this purpose, the SLM is estimated on a $K \times I$ binary matrix, where the entry cell $\mathrm{r}_{ki}$ is 1 if respondent *k* provided a response to item *i*, and 0 otherwise. The analysis results in a parameter $\beta_k$ for each respondent *k*, and a parameter $\gamma_i$ for each item *i*. The larger the value of $\beta_k$, the greater the propensity of *k* to respond to the items, whereas the larger the value of $\gamma_i$, the greater the tendency of *i* to be missed. According to the SLM, the greater the tendency of *i* to be missed and the lower the propensity of *k* to respond, the greater the probability that *k* misses *i*. The probability that *k* misses *i* equals .50 when $\beta_k = \gamma_i$, and exceeds .50 when $\beta_k < \gamma_i$.

Let us consider the following two sequences of responded "R" and missed "M" items, on a collection of 20 items ordered by increasing tendency to be missed:

$\beta_k$
RRRRRRRRRRRRRMMMMMMM
RRMRRMRRMRRMRRMRMMRR

$\longleftrightarrow$

Lower γ                Larger γ

The first sequence is too deterministic: $k$ systematically responded to all the items with $\gamma$ lower than $\beta_k$ and missed all the items with $\gamma$ larger than $\beta_k$. Conversely, the second sequence is too random: $k$ responded or missed the items regardless of the values of $\beta_k$ and $\gamma$. The outfit mean-square of the first sequence is expected to be smaller than 1, whereas that of the second sequence is expected to be larger than 1.

The outfit mean-squares can be converted into normally distributed $z$ values by means of the Wilson-Hilferty cube root transformation (Schulz, 2002; Wilson & Hilferty, 1931). This allows the testing of the null hypothesis that, on the whole, the response behavior displayed by the individual does not differ from that expected. An outfit $z$ value significantly smaller than 0 for a particular pattern of missingness suggests that the pattern might have been caused by a nonrandom mechanism.

## A SIMULATION STUDY

Several mixed data samples were simulated, each one consisting of complete response patterns, MCAR, MAR, and MNAR patterns. The study aims to investigate whether the five tests allow the distinction between the patterns of missingness deriving from a random mechanism (MCAR, MAR) and those deriving from a nonrandom mechanism (MNAR).

### Data Generation

The mixed data samples were generated using the following three-step procedure:

1. Simulation of the data sample with complete response patterns;
2. Random splitting of the data sample into four groups, and application of a different decimation mechanism (MCAR, MAR, or MNAR) to three of the four groups;
3. Merging of the four groups into the mixed data sample.

The procedure used for generating the mixed data samples is new, whereas the procedures used for simulating the three decimation mechanisms are well known in the literature (Finch, 2008; Hardouin, Conroy, & Sébille, 2011; Hohensinn & Kubinger, 2011; Holman & Glas, 2005).

The data samples with complete dichotomous (0, 1) response patterns were simulated using the SLM (Rasch, 1960). According to this model, the expected probability $p_{ki}$ that simulee $k$ provides a correct response to item $i$ is given by:

$$p_{ki} = \frac{\exp(\theta_k - \delta_i)}{1 + \exp(\theta_k - \delta_i)} \tag{8}$$

where $\theta_k$ is the latent trait value (ability) of simulee $k$, and $\delta_i$ is the scale value (difficulty) of item $i$. The expected probability $p_{ki}$ that $k$ provides a correct response to $i$ equals .50 when $\theta_k = \delta_i$, and exceeds .50 when $\theta_k > \delta_i$. All the data samples consisted of 100 items, whereas they differed in the number of simulees ($K$ = 100, 200, 500, 1,000, 5,000, and 10,000). The scale values $\delta$ were randomly drawn from a uniform distribution defined in the interval [−3.5; 3.5]. The latent trait values $\theta$ were randomly drawn from a normal distribution with $\mu = 0$ and $\sigma = 1.5$. A pseudorandom uniform value $u_{ki} \in [0, 1]$ was generated for each simulee $k$ and each item $i$. The response of $k$ to $i$ was correct (1) if $u_{ki} \leq p_{ki}$, and incorrect (0) otherwise.

Each data sample was randomly split into four groups, denoted as complete responses (CR), MCAR, MAR, and MNAR.

The response patterns of simulees in the CR group did not undergo any decimation mechanism. The response patterns of simulees in the MCAR group underwent a decimation mechanism that randomly replaced observed responses with missing responses. To simulate the different individual propensity to omit responses, the simulees in the MCAR group were randomly divided into three subgroups of equivalent size, and a missing response probability equal to .10, .30, or .50 was assigned to each subgroup. A pseudorandom uniform value [0, 1] was generated for each simulee and each item and, if it resulted to be lower than or equal to the assigned missing response probability, the item response was turned into a missing response. In the sequel, the probabilities .10, .30, and .50 will denote low, medium, and high missing propensity, respectively.

The response patterns of simulees in the MAR group underwent a decimation mechanism that made the missing response probability depending on the simulee's proportion of correct responses. Let $p_k$ be the proportion of correct responses given by simulee $k$ on the complete response pattern. The probability $m_k$ that $k$ provides a missing response was assumed to be exponentially decreasing with a higher $p_k$ according to the function: $m_k = mt \times \exp(-2 \times p_k)$. The value $mt$ represents the maximum of $m_k$, reached at $p_k = 0$. The simulees in the MAR group were randomly divided into three subgroups of equivalent size, to each of which a different $mt$ value was assigned. The $mt$ values were chosen to produce, in the three MAR subgroups, overall proportions of missing data that largely resembled those simulated in the three MCAR subgroups (i.e., .10, .30, .50). A pseudorandom uniform value [0, 1] was generated for each simulee and each item and, if it resulted to be lower than or equal to $m_k$, the item response was turned into a missing response.

The response patterns of simulees in the MNAR group underwent a decimation mechanism that made the missing response probability depending on both the simulee's latent trait $\theta_k$ and the item's scale value $\delta_i$. It is worth noting that, in the MCAR and MAR groups, the missing response probability did not depend on the item. Let $p_{ki}$ be the expected probability that simulee $k$ provides a correct response to item $i$, as computed by an application of Equation (8). The probability $m_{ki}$ that $k$ misses $i$ was assumed to be exponentially decreasing with a higher $p_{ki}$ according to the function: $m_{ki} = mt \times \exp(-2 \times p_{ki})$. The simulees in the MNAR group were randomly divided into three subgroups of equivalent size. The $mt$ values assigned to the three MNAR subgroups were chosen to produce overall proportions of missing data that largely resembled those simulated in the MCAR and MAR subgroups (i.e., .10, .30, .50). A pseudorandom uniform value [0, 1] was generated for each simulee and each item and, if it resulted to be lower than or equal to $m_{ki}$, the item response was turned into a missing response.

The four groups (CR, MCAR, MAR, MNAR) were then merged into the mixed data sample. A total of 12 scenarios was generated that differed for the total number of simulees ($K$ = 100, 200, 500, 1,000, 5,000, 10,000) and for the proportion of CR, MCAR, MAR, MNAR patterns in the mixed data sample — [CR = MCAR = MAR = MNAR = .25], [MNAR = .50, MCAR = MAR = .125, CR = .25]. In the scenarios [CR = MCAR = MAR = MNAR = .25], half of the missing patterns in the data sample derived from a random mechanism. In the scenarios [MNAR = .50, MCAR = MAR = .125, CR = .25], half of the missing patterns in the data sample derived from a nonrandom mechanism. For each scenario, 100 data samples were simulated from different values of parameter θ and δ (the MATLAB codes used for the simulations and the analyses are available upon request from the first author).

## Results

Figure 1 reports the results concerning the scenario [CR = MCAR = MAR = MNAR = .25] with 10,000 simulees. For each of the five tests, the bars represent the average proportions of statistically significant tests across the 100 simulated samples. The significance of the tests was evaluated by setting the Type-I error probability to .05. It is worth recalling that the null hypothesis specified that the mechanism underlying a particular pattern of missingness was random. The gray, white, and black bars respectively denote simulees with low, medium, and high missing propensity (overall proportions of missing data equal to .10, .30, and .50). The striped bars denote all simulees. MCAR, MAR, and MNAR patterns are in the upper, medium, and lower diagram, respectively. Comments to Figure 1 are in order.

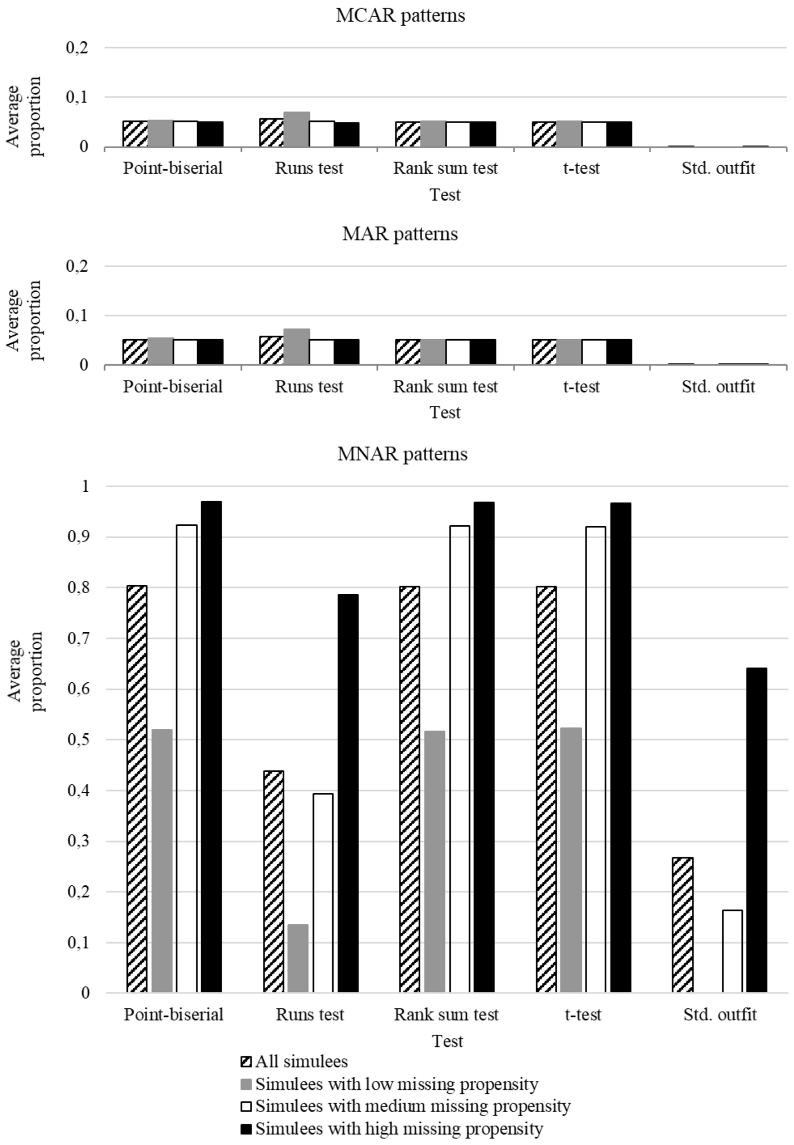

FIGURE 1
Average proportions of statistically significant tests (Type-I error = .05) across the 100 simulated samples. The null hypothesis specified that the mechanism underlying a particular pattern of missingness was random. Scenario [CR = MCAR = MAR = MNAR = .25] with 10,000 simulees.

Concerning the MCAR and MAR patterns, the proportion of tests statistically significant largely resembled the Type-I error probability of .05. This means that only 5% of these patterns were incorrectly identified as being caused by a nonrandom mechanism. Compared with the other tests, standardized outfit showed a stronger propensity in not rejecting the null hypothesis when this was true. The proportion of statistically significant tests did not vary with the missing propensity and, thus, with the number of missing responses in the patterns (on average, 10, 30, and 50 missing responses for simulees with low, medium, and high missing propensity, respectively).

The five tests differed in their propensity to reject the null hypothesis when this was false. Point-biserial correlation, rank sum test, and *t*-test correctly identified the largest proportions of MNAR patterns as caused by a nonrandom mechanism, followed by runs test and standardized outfit (lower diagram). Point-biserial correlation, rank sum test and *t*-test behaved in an almost identical way. The similar performance of rank sum and *t*-test is explained by considering that both tests are based on a measure (sum of ranks for the first test, average for the second test) of the difficulty of the items that were missed and responded by the simulee. Unlike rank sum and *t*-test, point-biserial correlation is based on a measure of the number of missing responses that the entire sample provided to the items that were missed and responded by the simulee. In our simulations, the number of missing responses vehiculated information about item difficulty through the MNAR patterns. In these patterns, the probability of observing a missing response increased with item difficulty.

The statistical power of each test increased with the missing propensity and, thus, with the number of missing responses in the patterns. For instance, point-biserial correlation, rank sum test, and *t*-test correctly identified 52% of MNAR patterns relative to low missing propensity simulees (about 10 missing responses), and 92% and 97% of MNAR patterns relative to medium and high missing propensity simulees (about 30 and 50 missing responses, respectively).

To investigate the statistical power of the tests in more detail, the patterns of missingness were divided into nine fractiles based on the number of missing responses (1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90; no patterns were observed with more than 90 missings). Results are depicted in Figure 2.

For point-biserial correlation, rank sum test, and *t*-test, the proportion of MNAR patterns that were correctly identified as being caused by a nonrandom mechanism approximately monotonically increases with the number of missing responses. Interestingly, these three statistics correctly identified a percentage of MNAR patterns as large as 95%, even when they were characterized by a relatively small number of missing responses (little more than 20 responses). A parabolic trend was observed for runs test and standardized outfit. With respect to runs test, this result is understood by considering that the observed number of runs (as such as its expected number) increases with the number of missing responses until the number of missing responses reaches half test length, whereas it starts decreasing with a number of missing responses exceeding half test length. Therefore, the performance of runs test with a few missing responses approaches that of the test with many missing responses. With respect to standardized outfit, it should be considered that extreme response patterns (in our case, all the items were responded or all the items were missed) always fit the SLM perfectly (Linacre, 2009). Therefore, the closer the proportion of missing responses to 0 or 1, the greater the probability that the standardized outfit approaches 0, so that a nonrandom mechanism can not be identified.

Figure 3 depicts the average proportion of MNAR patterns that, in each scenario, were correctly identified as caused by a nonrandom mechanism. Each graph pertains to a different test. Dashed and solid lines respectively represent the conditions [CR = MCAR = MAR = MNAR = .25] and [MNAR = .50, MCAR = MAR = .125, CR = .25]). The markers “×”, “□”, and “△” respectively represent low, medium, and high missing propensity. The size of the simulated samples is depicted in the *x*-axis. Comments to Figure 3 are in order.

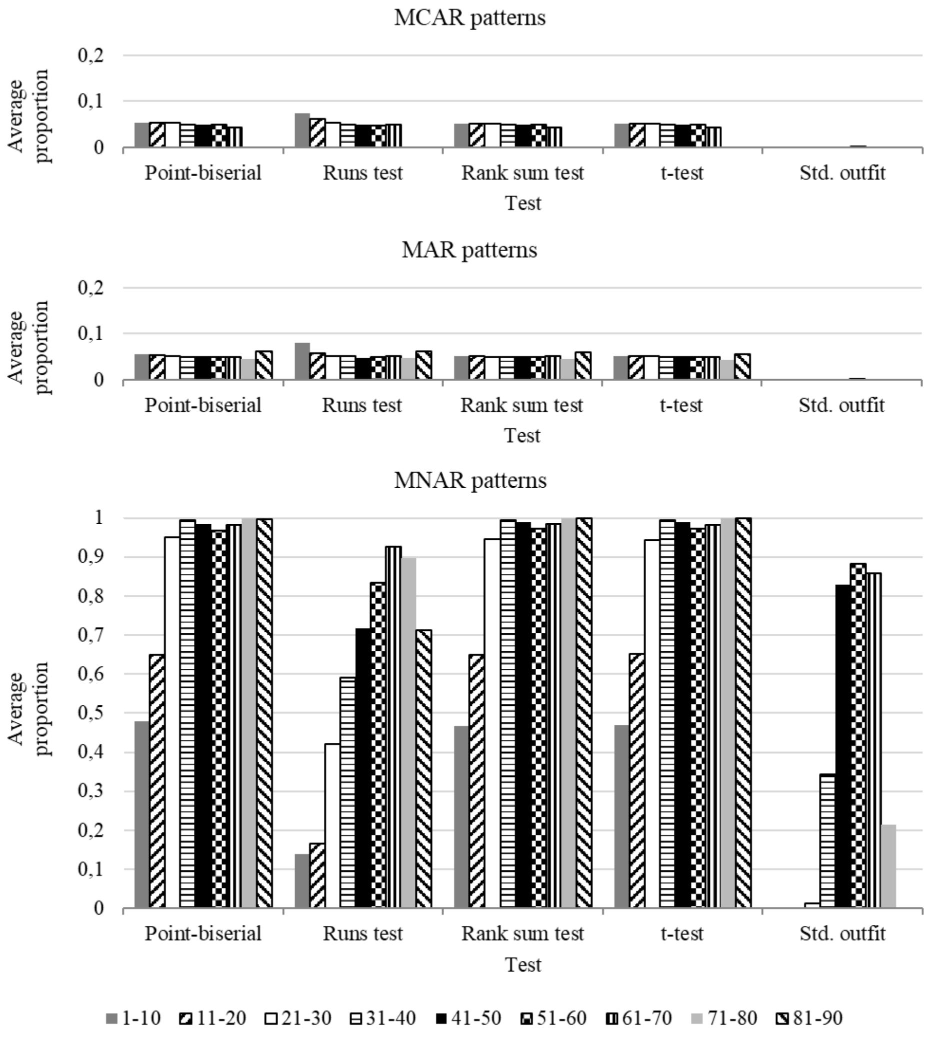

FIGURE 2
Average proportions of statistically significant tests (Type-I error = .05) across the 100 simulated samples.
The null hypothesis specified that the mechanism underlying a particular pattern of missingness
was random. Scenario [CR = MCAR = MAR = MNAR = .25] with 10,000 simulees.
Simulees divided into nine fractiles (the nine bars) based on the number of missing responses.

On the whole, the capability of each test to correctly identify the MNAR patterns as caused by a nonrandom mechanism decreased with sample size when the MNAR patterns do not represent the largest part of patterns in the sample (dashed lines). When the MNAR patterns represented half of the patterns (solid lines), the five tests were largely unaffected by sample size.

Point-biserial correlation, rank sum test, and *t*-test outperformed runs test and standardized outfit in all conditions of sample size and proportion of MCAR, MAR, and MNAR patterns in the sample. Standardized outfit turned out to be the worst test.

Regardless of sample size and proportions of random and nonrandom patterns, rank sum and *t*-test correctly identified more than 90% of MNAR patterns, provided that these patterns did not contain a negligible number of missing responses (lines marked by "□" and "△"). Point-biserial correlation performed worse than rank sum and *t*-test when there were not many missing responses resulting from a MNAR mechanism — condition [CR = MCAR = MAR = MNAR = .25] with 100 to 200 simulees and low to medium missing propensity. As mentioned above, rank sum and *t*-test are based on item difficulty, which enters point biserial correlation only indirectly through the number of missing responses. A larger amount of data is required for the number of missing responses to vehiculate proper information about item difficulty.

By applying Huisman's test (Huisman, 1999), the null hypothesis of random missingness was rejected for all the 100 mixed data samples simulated in each scenario, even when the sample size was not large (100 simulees) and the proportion of MAR and MCAR patterns was larger than the proportion of MNAR (.50 vs. .25).

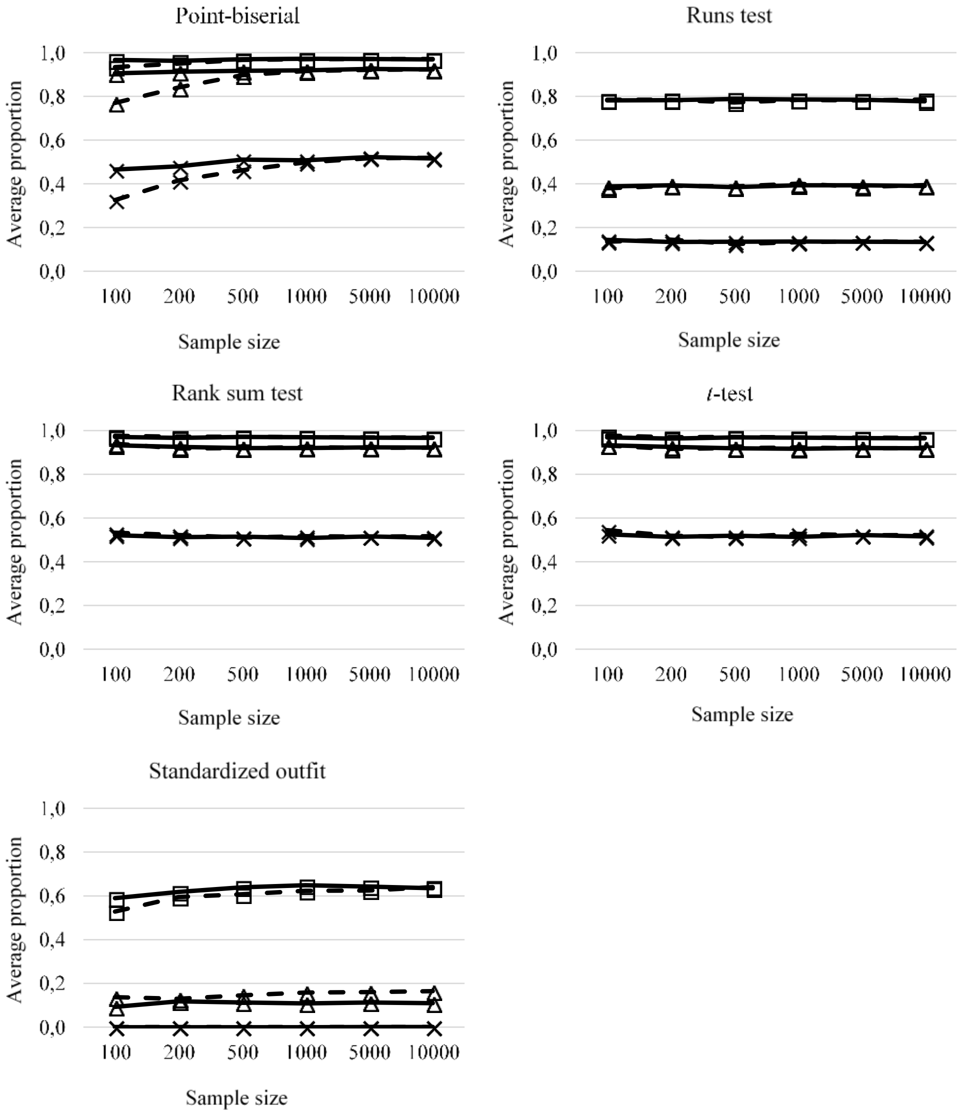

FIGURE 3
Average proportion of MNAR patterns correctly identified in each scenario.
Dashed and solid lines respectively represent the conditions [CR = MCAR = MAR = MNAR = .25] and [MNAR = .50, MCAR = MAR = .125, CR = .25]. The markers "×", "□", and "△" respectively represent low, medium, and high missing propensity. The size of the simulated samples is in the *x*-axis.

## AN EMPIRICAL DATA EXAMPLE

### Empirical Data

We used data from the information subtest of the National Intelligence Tests (Haggerty, Terman, Thorndike, Whipple, & Yerkes, 1920). An extensive description of the data can be found in Must and Must (2014), whereas the data are available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/23791. The information subset consists of 40 items scored as correct (1) and incorrect (0). Following the procedure described by Sijtsma and van der Ark (2003), we selected data from subjects who provided complete responses. Thus, the analyzed data sample consists of 311 respondents. From this data sample, 100 mixed data samples were created by applying Steps 2 and 3 of the procedure described in the simulation study. The proportion of CR, MCAR, MAR, and MNAR patterns was the same for all data samples (CR = MCAR = MAR = MNAR = 25).

### Results

Figure 4 depicts the results concerning the five tests, separately for MCAR, MAR, and MNAR patterns. The bars represent the average proportions of statistically significant tests (Type-I error probability of .05) across the 100 mixed data samples. The gray, striped, and white bars respectively denote patterns with a number of missings in the intervals 1-10, 11-20, 21-30 (no pattern was observed with more than 30 missings).

Concerning the MCAR and MAR patterns, the proportion of statistically significant tests resembled the Type-I error probability, and did not vary with the number of missing responses in the patterns. Compared with the other tests, standardized outfit showed a stronger propensity not to reject the null hypothesis when this was true. Analogous results were observed in the simulation study.

The five tests differed in their propensity to reject the null hypothesis when this was false. Rank sum and *t*-test correctly identified the largest proportions of MNAR patterns as caused by a nonrandom mechanism, followed by point-biserial correlation, runs test, and standardized outfit. We note that rank sum and *t*-test outperformed point-biserial correlation also in condition [CR = MCAR = MAR = MNAR = 25] of the simulation study, when the sample size resembled that of the empirical data sample (see Figure 3). The statistical power of each test increased with the number of missing responses in the patterns. For instance, rank sum test and *t*-test correctly identified 70% of MNAR patterns with 11-20 missing responses, and 95% of MNAR patterns with 21-30 missing responses. Analogous results were observed in the simulation study (see Figure 2). By applying Huisman's test (Huisman, 1999), the null hypothesis of random missingness was rejected for all the 100 mixed data samples.

## FINAL REMARKS

In the present work, missingness was investigated at the level of the individual respondent. The proposed approach represents a sharp departure from the traditional approach, in which missingness is investigated at the level of the entire group of respondents. Another novelty of the present work lies in the features of the data samples, each one consisting of patterns of missingness caused by different mechanisms.

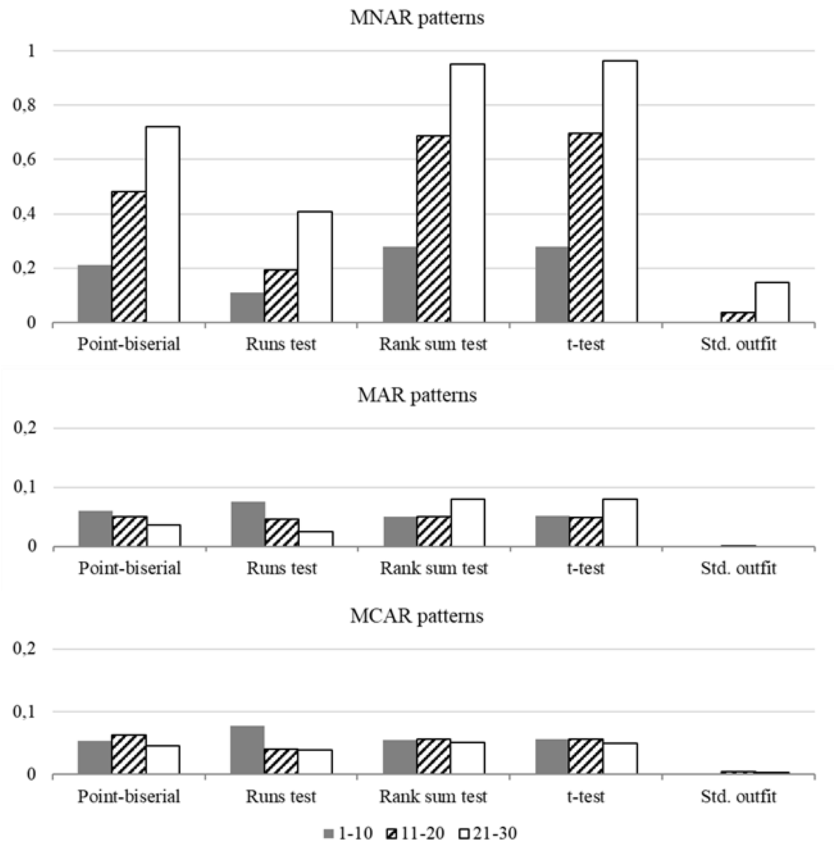

FIGURE 4
Average proportions of statistically significant tests (Type-I error = .05) across the 100 mixed data samples. The null hypothesis specified that the mechanism underlying a particular pattern of missingness was random.

In other works (Hohensinn & Kubinger, 2011; Holman & Glas, 2005; Sijtsma & van der Ark, 2003), a unique missing mechanism underlies the entire data sample. This condition might not reflect what happens in real situations.

Among the considered tests, Wilcoxon rank sum test and *t*-test resulted the best ones for correctly identifying MNAR patterns under different conditions of sample size, proportions of MCAR, MAR, and MNAR patterns in the data sample, and number of missing responses in the patterns. Both tests are based on a measure of the difficulty of the items that were missed and responded by the individual. Even when the sample size was not large, rank sum test and *t*-test allowed the identification of a very large percentage of MNAR patterns (more than 90% in the present study), provided that the number of missing responses in each pattern was not negligible. These results suggest that rank sum test and *t*-test can be used to infer if the particular pattern of missingness displayed by a respondent has been caused by a random or nonrandom mechanism. Such an inference can not be made using traditional methods, such as Huisman's test (Huisman, 1999), which are meant to analyze missingness at the level of the entire data sample. We note in passing that, whenever it is necessary to obtain an overall judgement about missingness at the sample level, results of the individual tests can be combined into a unique overall statistic (see, e.g., Fisher, 1932).

It is worth noting that, once there is adequate information about item difficulty and propensity to be missed (e.g., derived from the analysis of data samples already available), the proposed tests can be used to evaluate the pattern of missingness displayed by a new individual respondent without having to collect data on an entire sample of respondents.

A variety of methods for dealing with missing data have been proposed in the literature. These range from imputation methods for replacing the missing data with substituted values (Hardouin et al., 2011; Sijtsma & van der Ark, 2003) to formal models for analyzing incomplete data samples (Anselmi, Robusto, Stefanutti, & de Chiusole, 2016; de Chiusole, Stefanutti, Anselmi, & Robusto, 2015; Holman & Glas, 2005). The choice of the imputation method or the formal model requires an assumption about the mechanism that caused missingness. For instance, two highly recommended imputation methods, multiple imputation (Rubin, 1987) and maximum likelihood (Little & Rubin, 1987), assume that the data are MCAR or, at least, MAR. Biases might result from an application of these methods to MNAR data (Finch, 2008). Similarly, some formal models for the analysis of incomplete data samples assume random missingness. Biases in the estimation of person and item parameters might derive from applying these models to MNAR data (de Chiusole et al., 2015). If a set of data is available, Wilcoxon rank test and *t*-test can be used to distinguish the patterns of missingness that, with some certainty, are caused by a nonrandom mechanism from those that are caused by a random mechanism. Then, appropriate analysis methods can be applied to the two collections.

## References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*, 1012-1028. doi:10.1111/j.1741-3737.2005.00191.x

Anselmi, P., Robusto, E., Stefanutti, L., & de Chiusole, D. (2016). An upgrading procedure for adaptive assessment of knowledge. *Psychometrika*, *81*, 461-482. doi:10.1007/s11336-016-9498-9

Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes*, *13*, 16. doi:10.1186/s12955-014-0197-x

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.

Colledani, D., Robusto, E., & Anselmi, P. (2018). Development of a new abbreviated form of the Junior Eysenck Personality Questionnaire-Revised. *Personality and Individual Differences*, *120*, 159-165. doi:10.1016/j.paid.2017.08.037

Da Dalt, L., Anselmi, P., Bressan, S., Carraro, S., Baraldi, E., Robusto, E., & Perilongo, G. (2013). A short questionnaire to assess pediatric resident’s competencies: The validation process. *Italian Journal of Pediatrics*, *39*, 41. doi:10.1186/1824-7288-39-41

Da Dalt, L., Anselmi, P., Furlan, S., Carraro, S., Baraldi, E., Robusto, E., & Perilongo, G. (2015). Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. *Italian Journal of Pediatrics*, *41*, 2. doi:10.1186/s13052-014-0106-2

de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2015). Modeling missing data in knowledge space theory. *Psychological Methods*, *20*, 506-522. doi:10.1037/met0000050

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *4*5, 225-245. doi:10.1111/j.1745-3984.2008.00062.x

Fisher, R. A. (1932). *Statistical methods for research workers* (5th ed.). Edinburgh, Scotland: Oliver & Boyd.

Gibbons, J. D. (1985). *Nonparametric statistical inference* (2nd ed.). New York, NY: Marcel Dekker, Inc.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston, MA: Allyn & Bacon.

Haggerty, M., Terman, L., Thorndike, E., Whipple, G., & Yerkes, R. (1920). *National Intelligence Tests. Manual of Directions. For Use with Scale A, Form 1 and Scale B, Form 1*. New York, NY: World Book Company.

Hardouin, J-B., Conroy, R., & Sébille, V. (2011). Imputation by mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Medical Research Methodology, 11*, 105.
doi:10.1186/1471-2288-11-105

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, *53*(3), 380-393.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York, NY: Wiley.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-17.
doi:10.1348/000711005X47168

Huisman, J. M. E. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

Linacre, J. M. (2009). *Winsteps (Version 3.68.0)* [Computer software]. Chicago, IL: Winsteps.com.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477-482.
doi:10.1007/BF02293689

Matters, G., & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement*, *63*, 239-256.
doi:10.1177/0013164402250988

Must, O., & Must, A. (2014). Data from "Changes in test-taking patterns over time" concerning the Flynn Effect in Estonia. *Journal of Open Psychology Data*, *2*, e2.
doi:10.5334/jopd.ab

Osborne, J. W. (2013). *Best practices in data cleaning*. Los Angeles, CA: Sage Publications, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: Danish Institute for Educational Research.

Rubin, D. B. (1976). Inference and missing data. *Biometrika Trust*, *63*, 581-592.
doi:10.1093/biomet/63.3.581

Rubin, D. B. (1987*). Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.
doi:10.2307/3002019

Schulz, M. (2002). Standardization of mean-squares. *Rasch Measurement Transactions*, *16*(2), 879.

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*, 505-528.
doi:10.1207/s15327906mbr3804_4

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, *169*, 1133-1139.
doi:10.1093/aje/kwp026

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291-307.
doi:10.1177/1073191110374797

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, *11*, 147-162.
doi:10.1214/aoms/1177731909

Weeks J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671-701.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*, 28-35.
doi:10.2307/2332510

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, *17*(12), 684-688.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D, & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*, 18.
doi:10.1186/s41155-016-0040-x