Running title: Effects of 'Does Not Apply'

The Effects of 'Does Not Apply' on Measurement of Temperament with the Infant Behavior Questionnaire-Revised: A Cautionary Tale for Very Young Infants

Gerald F. Giesbrecht, Deborah Dewey, and the APrON Study Team

University of Calgary

Author Note

Gerald Giesbrecht, Department of Paediatrics, Faculty of Medicine, University of Calgary, Alberta Children's Hospital Research Institute for Child and Maternal Health, Calgary, Alberta Canada.

Deborah Dewey, Departments of Paediatrics and Community Health Sciences, Faculty of Medicine, University of Calgary, Alberta Children's Hospital Research Institute for Child and Maternal Health, Calgary, Alberta Canada. Email: dmdewey@ucalgary.ca

Correspondence concerning this article should be addressed to Gerald Giesbrecht, Behavioural Research Unit, Alberta Children's Hospital, 2888 Shaganappi Trail NW, Calgary AB, T3B 6A8, Canada. E-mail: ggiesbre@ucalgary.ca. Phone: (403) 955-2793.

1.0 Introduction

Missing data within multi-item scales is common within developmental and social science research, but surprisingly little guidance is available for how to deal with the problem. The paucity of attention to this topic is in stark contrast to the vast (and rapidly growing) literature addressing unit nonresponse - when the entire scale or variable is missing for an individual [1]. In this paper, we describe a missing data technique for multi-item scales that is commonly used but seldom recognized as method for dealing with missing data. Its lack of recognition may be due in part to the fact that the method does not even have a name; we propose the term "scale mean substitution". Our goal is to raise awareness of this missing data technique, illustrate its effects on data quality, and suggest ways to maintain data quality on multi-item scales.

1.1. Scale mean substitution is not (exactly) the same as person mean imputation

Scale mean substitution (SMS) is one of only a few missing data techniques that specifically address item nonresponse on multi-item scales. The procedure is a simple two-step process: 1) calculate a mean score for each person using all items with a valid score and ignore all missing items, and 2) use this mean of the available items as the substitute for the scale mean that would have been obtained if the person had responded to all items (thus our use of the term substitution). This procedure assumes that it is more reasonable to estimate a person's standing on a construct from the available data on a multi-item scale than to report a missing value for the entire scale – an assumption that is reasonable when all items on the multi-item scale are at least moderately correlated and measure a single construct. The primary reason for assessing a construct using multiple items is to increase the reliability of its estimation by triangulating from the individual items [2]. When individual items are missing all is not lost; the other items in the

scale should be informative of the person’s standing on the construct. However, accepting a score based upon a reduced set of items raises conceptual and statistical difficulties. Not only is the scale less reliable because fewer items are used to calculate the score, but also the definition of the scale can no longer be defined *a priori* because it now depends on the particular pattern and rates of item response, which may vary from person to person. Such variation “violates the principle that an estimand be a well-defined aspect of a population, not an artefact of a specific data set” (p. 158) [3]. For developmentalists, the problem of scale definition is particularly bothersome if the rates of missing data change over time. To the extent that some items on a scale are neglected at one age but endorsed at another age, the underlying construct assessed at the two time points may not be equivalent and estimates of individual change over time will be degraded.

SMS should not be confused with a highly similar approach known as person mean imputation (PMI) [3]. PMI interposes an additional step between steps one and two above. Specifically, the person’s mean, based on available items as is done in the SMS approach, is imputed for all missing items, resulting in a complete data set. The scale means are then calculated using the complete data. We note that both SMS and PMI will produce the same means and variances because imputing a mean does not add or remove any information regarding an individual’s standing on the construct measured by the scale. However, the assumptions underlying these two methods are different and the effects on the validity and reliability of the resulting data are distinct. The PMI approach assumes that the person’s actual score on a missing item can be best estimated by imputing the person’s mean. In contrast, the SMS approach is agnostic about the person’s standing on the missing item and relies instead on the notion that because all items assess a single construct the remaining items can be used to

estimate the person’s standing on the construct. One way in which these two approaches differentially affect reliability estimates is that individuals with missing data will be eliminated from any calculation of scale reliability using the SMS approach but they will be included when using the PMI method. As described in our analysis below, this difference has a dramatic effect on estimates of scale reliability and, by extension, on validity.

1.2. The Current Study

The purpose of the current study was to illustrate the effects of SMS on the quality of data derived from a commonly used multi-item temperament scale, the Infant Behaviour Questionnaire-Revised (IBQ-R) [4], and to evaluate the effects of various missing data strategies. The IBQ-R has 14 multi-item scales that assess different aspects of temperament in infants between 3- and 12-months of age. Our interest in missing data on the IBQ-R derives from its ‘does not apply’ response option that does not yield a scorable response. These data are not technically ‘missing’ because the parent actually did provide a response, but they are missing in the sense that they cannot be included in the individual’s score. The IBQ-R uses the SMS procedure to calculate scale scores and the effects of this procedure on infant temperament scores have not previously been reported.

Our analysis describes the nature, extent, and pattern of missing items on the short form of the IBQ-R at 3- and 6-months of age in a community sample of 458 infants. We then demonstrate the effects of missing items on the internal consistency, scale means, and longitudinal stability of the scales using the SMS procedure and compare them to results obtained using alternate missing data strategies. Finally, we describe correlates of missing items. In the discussion, we suggest general guidelines for using SMS as a missing data technique in developmental and social science research and make specific recommendations for the IBQ-R.

2.0 Material and Methods

2.1. Infant Behavior Questionnaire – Revised

The original Infant Behavior Questionnaire [5] was developed to study individual differences in infant temperament that could be observed and reported by parents. Since its publication in 1981, it has generated a great deal of interest, and with it, scrutiny of its psychometric properties, which resulted in a revised version [4]. The short form of the IBQ-R [6] is a 91 item measure composed of 14 scales: Activity Level, Distress to Limitations, Fear, Duration of Orienting, Smiling and Laughter, High Intensity Pleasure, Soothability, Falling Reactivity/Rate of Recovery from Distress, Cuddliness, Perceptual Sensitivity, Sadness, Approach, and Vocal Reactivity. Parents report their observations of specific infant behaviours in the previous two weeks using a 7-point Likert scale that ranges from ‘never’ to ‘always’. The short from of the IBQ-R is reported to have similar validity and reliability compared to its longer version [6].

Missing items on the IBQ-R are of two types: items that are left blank and items that are marked as ‘does not apply’ (NA). Blank items occur for a variety of (ultimately unknown) reasons. In contrast, NA items occur when the parent is unable to provide a scorable response (i.e., a response on the scale of ‘never’ to ‘always’) because the behavior and situation in question did not occur in the previous two weeks. The scoring procedure for the IBQ-R does not differentiate between items that are left blank and items that are marked as NA – in both cases these items are excluded from the calculation of the scale score.

2.2. Procedure

The short form of the IBQ-R and several other infant questionnaires were mailed to participants two weeks prior to target assessment dates of 3-months (T1) and 6-months of age

(T2). Participants were asked to complete the questionnaire at home and return it at their scheduled 3-month study visit (T1) or return it using a prepaid mailing envelope (T2).

Our instructions for the IBQ-R included a minor modification to the standard IBQ-R instructions to ensure that parents were clear on the use of the ‘does not apply’ and ‘never’ response options. The existing instructions stated:

> The “does not apply” (NA) column is used when you did not see the baby in the situation described during the last week. For example, if the situation mentions the baby having to wait for food or liquids and there was no time during the last week when the baby had to wait, circle the (NA) column. “Does not apply” is different from “never”. “Never” is used when you saw the baby in the situation but the baby never engaged in the behavior listed during the last week. For example, if the baby did have to wait for food or liquids at least once but never cried loudly while waiting, circle the “never” column.

In consultation with Sam Putnam (personal communication November 30, 2009), we appended the following clarification to the above instructions: “Some of the items in this questionnaire are beyond the abilities of many 3-month olds or 6-month olds. Please use the “does not apply” (NA) option for behaviours that your infant does not yet do.”

2.3. Participants

The Alberta Pregnancy Outcomes and Nutrition study [7] is a prospective cohort study of pregnant women and their infants designed to examine the effects of prenatal exposures on infant development. Women were recruited during pregnancy between June 2009 and July 2012 from two metropolitan areas through advertisements in local media and by stationing research assistants in waiting rooms of high volume maternity care and ultrasound clinics. Ethics approval for this study was obtained from the Health Research Ethics Boards at the University of Alberta

and the University of Calgary in Alberta, Canada. Written consent was obtained from all women prior to enrolment in the study.

IBQ-R data were available for 458 3-month-old infants and 391 6-month-old infants. Six infants were twin pairs – in order to maintain assumptions of data independence, we randomly eliminated one of the twins from each pair. An additional 23 infants born at less than 37 weeks of gestation were excluded in order to avoid the possibility that some data were missing because of developmental delays associated with preterm birth [8]. Infants for whom the IBQ-R was completed more than 4 weeks prior to or after the target dates of 3- and 6-months of age were also excluded from the analyses. This resulted in an additional 57 infants who were excluded at T1 and 66 infants who were excluded at T2.

The final sample of 401 infants at T1 had a mean age of age 2.7 months (range 2.0 - 4.0); 44 % were female. At T2, 325 infants had a mean age 6.2 months (range 5.0 - 7.0); 45% were female. The mothers were primarily White (85.7 %), with annual household income above $40,000 (93.1%), married or living common law (96.8%), and well educated (91.4% had some training beyond high school). The mean age of mothers at T1 was 31.2 years (SD = 4.1; range = 20.3 – 43.2) and 56.7% of infants were a first child.

2.4. Statistical Analysis

All analyses were conducted using SPSS 20.0. We calculated scale means using the SPSS syntax included in the IBQ-R scoring manual. As per the standard scoring procedure, we used a SMS procedure; that is, we calculated a mean of available items (after reverse scoring items, as appropriate).

To highlight the effects of NA responses on each of the IBQ-R scales, we first identified infants with no NA responses and those who had two or more NA responses for a particular

scale. We selected infants with two or more items marked as NA for the following reasons: i) it is approximately equivalent to 30% missing data and there is general agreement that this proportion of missing data requires a missing data strategy, and ii) we wanted to create a comparison that was meaningfully based upon the rates of missing data actually observed in the data. For each scale, we created a group with no NA responses (No NA group) and a group with at least 30% of items marked as NA (NA ≥ 30% group). Sample sizes for these groups varied by scale and assessment occasion. To avoid spurious differences, when group size dropped below 10% of the total sample, we did not report a result. Our rationale for comparing these groups was to determine the extent to which the presence of NA responses might be responsible for inter-individual differences in temperament scores. To the extent that there are group differences in the observed reliability, means and longitudinal stability of temperament that can be attributed to the presence of NA responses, such group differences could add noise to the data and interfere with assessment of infant temperament using the IBQ-R.

To evaluate the potential to improve upon the SMS procedure, we compared the results obtained from the SMS produce to those obtained by four alternate missing data strategies. Our goal was to determine if these alternate missing data strategies could eliminate differences in reliability and means between the No NA and the NA ≥ 30% groups. The first strategy was PMI. As described above, PMI imputes the person's scale mean obtained from the remaining items for any missing items. Although this procedure is known to attenuate variance and may produce inconsistent bias to parameter estimates (sometimes leading to Type 1 and other times leading to Type 2 errors) [9], it appears to produce reasonably unbiased parameter estimates when the overall rates of missing items are below 20% [10, 11]. Our selection of the second strategy, which we refer to as NA = never, was based on the notion that the best way to deal with missing

data is to not have it in the first place. We reasoned that it may be possible to eliminate the NA response option from the IBQ-R and instead instruct parents to use the ‘never’ response option in its place. For this analysis, we replaced all NA responses with a value of 1 corresponding to ‘never’. The third missing data strategy is expectation maximization (EM). The EM strategy is based on a maximum likelihood algorithm that uses an iterative procedure to estimate missing values using all available data, not just complete cases. EM is known to produce unbiased parameter estimates; however, it has a tendency to reduce standard errors and should be used with caution for hypothesis testing [12]. We used the EM option within the Missing Values Analysis package from SPSS to impute a single complete data set. Our final strategy was multiple imputation (MI), as implemented in SPSS, to impute values for 10 complete data sets. MI first makes multiple copies of the data set and then independently imputes missing values for each data set using a stochastic method to add variation from one imputation to another. Data analyses are then run separately for each imputed data set and the results are pooled to provide a single estimate. MI solves the problem of underestimating uncertainty that is a limitation for EM [3] and has been previously applied to the problem of missing data on multi-item scales [13].

3.0 Results

3.1 Frequency of Missing and NA Items

Missing data due to items left blank were rare (< 0.2% at both ages), whereas missing data due to selection of the NA option were high at T1 (22.4%) but much lower at T2 (7.4%). The mean number of NA items per infant at T1 was 20.4 (SD = 12.1; range 0 - 59) and 6.7 (SD = 5.9; range 0 - 31) at T2. Examination of the NA responses on individual scales (see Table 1) suggested that NA responses tended to cluster within some scales. The mean number of NA items exceeded 50% in the Perceptual Sensitivity and Approach scales at T1 and two additional

scales, Duration of Orienting and High Intensity Pleasure, had NA items exceeding 30%. None of the scales at T2 had NA rates that exceeded 50% and only the Perceptual Sensitivity scale exceeded 30% NA responses.

3.2. Effects of NA responses on Internal Consistency

As seen in Tables 2a and 2b, the proportion of cases with complete data (that is, the No NA group) on the individual IBQ-R scales ranged from 15% to 79% at T1 and 27% to 93% at T2. The alphas calculated using the SMS procedure were adequate ($\geq .70$), with the exception of the Fear scale ($\alpha = .61$) at T1 and the Sadness and Approach scales ($\alpha = .68$ for each) at T2. However, it should be noted that because alpha requires complete data for every item on a scale, more than 50% of the sample was eliminated for nine of these scales at T1 and one scale at T2. For the purpose of calculating alpha, the SMS procedure is equivalent to a listwise deletion strategy in which any individual with missing data is excluded.

Tables 2a and 2b also display the alphas produced by the missing data strategies. Alphas could not be calculated for infants with NA responses using the SMS procedure because it does not actually impute values and complete data is required to run the analysis. To obtain an estimate of the actual alphas using the SMS procedure in the presence of NA responses, we first eliminated *items* (as opposed to individuals, as is done in listwise deletion) for which more than 30% of the sample recorded an NA response. We then calculated the alphas based upon these shortened scales. Not surprisingly, the reduction in items attenuated the alphas, which ranged from .30 to .86 at T1 and .52 to .84 at T2. Of note, alphas were below .70 on seven scales at T1 and three scales at T2. An alpha could not be calculated for the Perceptual Sensitivity scale at T1 because every person in the study had at least one missing item.

As shown in the four rightmost columns of Tables 2a and 2b, there was considerable variability in the Cronbach's alphas produced by the alternate missing data strategies. The PMI procedure resulted in alphas that were substantially higher than all other methods. The NA = never strategy had the opposite effect; alphas were substantially lower compared to all other methods. The EM method produced modest estimates that were at or above .70 for most scales at T1 and all of the scales at T2. The MI procedure resulted in alphas that were lower than the EM method and fell below .70 on many of the scales.

3.3. Effects of NA Responses on Scale Means

There were large effects of NA responses on the scale means, especially at T1 when the rates of missing data were high. The SMS and PMI methods (which yielded equivalent results for this analysis) resulted in means that were generally higher among infants in the NA > 30% group compared to the No NA group (see Tables 3a and 3b). Means were significantly higher ($p < .05$) on four scales at T1 and two scales at T2. The NA = never approach produced means that were significantly lower ($p < .01$) compared to cases with complete data for every scale at both T1 and T2. The EM method resulted in means that were equivalent to the No NA group, with the exception of the Vocal Reactivity scale at T1 and the Approach scale at T2. Similarly, the MI approach generated means that differed ($p < .05$) from the No NA group on only 2 scales at each time point.

3.4. Effects of NA responses on Longitudinal Continuity of Temperament

Longitudinal continuity between the T1 to the T2 assessments are reported in Table 4. The average correlation between T1 and T2 assessments for the No NA group was $r = .46$. All missing data strategies for the NA ≥ 30% group resulted in lower correlations compared to those obtained from the infants in the No NA group. All methods produced very low correlations for

Falling Reactivity and Cuddliness. In contrast, on the Soothability scale, all missing data methods produced higher correlations relative to the No NA group. Across all scales, the NA = never and the MI strategies produced the lowest correlations between T1 and T2, mean r = .31 for both methods. The SMS and PMI approaches (which yielded equivalent results) had an average correlation of r = .34. The EM strategy had the highest average correlation at r = .42.

3.5. Correlates of NA Response

Given that scale means in the NA ≥ 30% group were elevated relative to the No NA group when the usual (SMS) scoring procedures was used, we wondered if the reason for this difference might be related to differences in the use of the 'never' response option. To test this question, we compared the use of never (1) ratings relative to all other ratings between 2 and 7. Across all scales, the proportion of never responses was 21% in the No NA group and 8% in the NA ≥ 30% group. This suggests that that use of the NA response option tends to increase mean scores by decreasing the number of 'never' responses.

In order to more fully understand the context that gives rise to NA responses, we examined associations between their occurrence and demographic and personal characteristics of the mother. For this analysis, we used the total number of NA responses for each mother. At both T1 and T2, number of NA responses was not related to marital status, amount of education, household income, ethnicity, maternal age, infant sex, or concurrent symptoms of depression or anxiety (assessed via the Edinburgh Postnatal Depression Scale [14] and the Symptom Checklist 90-R Anxiety scale [15]). In contrast, number of NA responses was correlated with number of children, $r(395) = .23$, $p < .001$ at T1 and $r(320) = .12$, $p < .05$ at T2. At T1, but not at T2, infant age was correlated with number of NA responses, $r(401) = -.23$, $p < .001$. These findings indicate

that mothers of younger infants produced more NA responses at T1, and at both T1 and T2 mothers with more children selected the NA response more often.

4.0 Discussion

4.1. Summary of Results

This study examined the occurrence and effects of missing data on the short form of the IBQ-R. Surprisingly few items (< 0.2%) were missing because they were left blank, whereas missing data resulting from NA responses was considerable for 3-month-old infants (22%) but significantly lower for 6-month-old infants (7%). Some IBQ-R scales had more missing data than others, suggesting that parents have relatively more difficulty reporting on some aspects of temperament, especially in early infancy. There were noticeable reductions in internal consistency associated with missing data, but this problem was limited primarily to assessment of 3-month-old infants. There was clear evidence that the SMS procedure tends to inflate mean temperament scores in the presence of NA responses. In addition, the longitudinal stability of some, but not all, scales was decreased when data was missing. Finally, both infant age and number of children were associated with the prevalence of NA responses.

4.2. General Implications for Multi-Item Scales

The analysis we conducted focused on an infant temperament instrument, but the results are applicable to other multi-item measures with missing data. In the following discussion, we highlight reliability and validity as issues that are of general relevance to researchers with missing data on multi-item scales.

The number of items used to measure a construct is directly related to its reliability [16]. When a sizable portion of the sample has missing data for one or more items, reliability will be attenuated. Reliability estimates reported in the research literature are based on complete data,

and as was the case for the IBQ-R, estimates based on the SMS procedure overestimate the actual performance of a scale when items are missing [2]. Accordingly, in the absence of an adequate missing data strategy, the internal reliability estimates reported for multi-item scales should be taken as the upper boundary of potential reliability with the actual reliability somewhat lower and dependent on the rate of missing data.

Reliability also places an upper limit on the validity of a scale [17]. When some items are missing, not only is the estimate of a person’s standing on the construct less reliable, it is also somewhat inconsistent with its theoretical definition [2]. The problem is one of construct coverage – the extent to which the individual items actually assess the target construct. If some items are consistently missing for all individuals, then part of the construct is systematically missing; however, if individuals differ with regard to their pattern of missing items, then the nature of the construct will be somewhat different for each person. In either case the construct is poorly defined and the “scores obtained provide an inaccurate as well as imprecise reflection of an individual’s standing on the construct” (p. 23) [2].

In the context of longitudinal research, differential patterns of missing items from one occasion to another may not only contribute to confusion about the definition of the construct on each occasion but may also lead to poor stability of the construct, especially when important aspects of the construct are missing at one point but not the other. Such considerations are important for the IBQ-R, which is often used to track changes in temperament over time.

4.3. Implications for the IBQ-R

Beyond the general implications for multi-item scales, several issues specific to the IBQ-R require further elaboration. Perhaps the most important observation about missing data on this measure is that the pattern of missing data speaks to the difficulties parents have when reporting

on the behaviours of their infants, especially very young infants. Almost all (> 99%) of the missing data in the current study resulted from NA responses. In other words, parents *did* respond to the item, and it is therefore not technically missing, but the response was not scorable. Although NA responses cannot help us understand individual differences in infant temperament, they do tell us something about the way that parents interact with the IBQ-R. The fact that rates of NA response were three times as high when infants were 3-months old compared to 6-months old indicates that the difficulties parents have in rating their infant's behaviour at an earlier age are largely resolved with subsequent development and opportunities for observation. It is also instructive to note that parents who did *not* use NA responses used never responses three times as often as parents who *did* use NA responses. The reasons that some parents do use NA responses while others do not requires further investigation. We found evidence suggesting that mothers who were reporting on a first child had a significantly lower NA response rate, suggesting that the way mothers report infant behavior on the IBQ-R is influenced by parenting experience. However, the importance of parenting experience on patterns of NA response seems to diminish over the second three months of the infant's life, as is shown by the decrease in the correlation between NA responses and parity from T1 to T2.

4.4. Recommendations for Scoring IBQ-R Scales

4.4.1. Delay assessment until 6 months

The majority of NA responses occur with reference to very young infants. Perhaps the most obvious solution is to delay assessment until infants are at least 6 months of age. Overall rates of NA items were less than 7% at 6 months and problems related to reduced reliability because of scale shortening were detectable but not a major concern, with the exception of the Perceptual Sensitivity scale, which continued to garner many NA responses at 6 months. We are

aware that assessing early infant temperament (before 6-months) has become an important area of research and some researchers, including our own group, find the recommendation to wait until 6 months to be unsatisfying. Our findings should at least give pause to researchers interested in very young children and encourage them to examine the effects of NA responses on their data.

4.4.2. Eliminate some scales at younger ages

Given that NA items tend to accumulate within some scales and not others, an alternate approach might be to include only those scales that have lower rates of NA items when assessing the youngest infants. In our analysis, the Duration of Orienting, High Intensity Pleasure, Perceptual Sensitivity, and Approach scales had rates of NA response that exceeded 30% at 3 months. The high rate of NA response on these scales strains their credibility when used to assess very young infants. The difficulty with these scales appears to resolve by the time infants reach 6-months of age, with the exception of the Perceptual Sensitivity scale. Thus one approach would be to use only those scales with lower rates of NA responses at 3 months.

4.4.3. Generate a version specifically for 3 months of age

It may be possible to create a version of the IBQ-R specifically for use at 3-months of age. We note that Gartstein and Rothbarth [4] considered this approach in their revision of the Infant Behavior Questionnaire but elected to retain a common item pool across the first year because rates of NA items did not differ between younger and older infants. The difference between their finding and the current findings deserve consideration. The most likely explanation for why Gartstein and Rothbart did not find age-related changes in NA responses is that they analysed their data using age brackets that may have obscured these differences. They collected data from infants who were 3- to 12-months old and then grouped them by age: 3-6 months, 6-9

months, and 9-12 months. Our data shows a strong decrease in the number of NA items from 3- to 6-months of age. Furthermore, we excluded infants who were more than 4 months old at T1 or younger than 5 months old at T2 in order to make a clear distinction between the 3- and 6-month old infants. Unfortunately, Gartstein and Rothbart did not report the age composition for their 3-6 month group. If the mean age of their sample was closer to 6 months than to 3 months, then the findings for our sample at 6 months of age would be consistent with their reported findings.

4.4.4. Use a longer version of the IBQ-R

One of the primary limitations of the current analysis is that it is based upon the short form (91 items) of the IBQ-R. Many of the problems related to NA responses that we have identified, such as scale reliability and inflation of means, may be attenuated with the standard version (191 items) of the IBQ-R. Although participant burden sometimes dictates that a shorter form of the scale must be used, this choice should be balanced with missing data considerations, especially in early infancy when the rate of NA responses is relatively high. Note, however, that a longer version of the IBQ-R does not resolve theoretical issues related to the definition of the underlying construct as it would continue to be assessed by a different set of items for each individual.

4.4.5. Use a modern missing data technique

Our evaluation of the SMS, PMI and NA = never strategies suggest that these approaches may influence the results of analyses based upon them. The SMS and PMI approaches yielded significantly higher means and substantially lower longitudinal stability compared to infants without NA responses, especially at 3 months. This should give pause to researchers interested in development, and especially those interested in the causes of intra-individual change over time, because some intra-individual changes in temperament may be an artifact of NA responses.

Consider an infant with a high number of NA responses on the Fear scale at 3 months. All things being equal, this infant is likely to have an above average score on the Fear scale simply by virtue of missing data. If the infant is assessed again later and the number of NA responses is reduced, this infant's Fear scale is likely to be much closer to the mean. Note that this shift in rank order does not take into account development – it is purely an artifact of scoring procedure. To the extent that NA responses affect the rank order of many infants within the sample, the SMS and PMI approaches could alter estimates of longitudinal stability.

The NA = never has the potential to cause the same difficulty as noted for the SMS and PMI approaches. The only difference is that the mean score for an infant with a high number of NA responses at 3 months would be lower than average. The NA = never approach has the appeal of simplicity but its performance was very poor. Internal consistency was very poor and means were statistically lower for every scale at both 3 and 6 months. We note, however, that our analysis of this approach is limited by the fact that parents were instructed to make a distinction between the never and NA response options. It remains to be seen whether this strategy is viable if the NA response option were removed from the instructions. Our view is that removing the NA option would be unwise as it may reduce the face validity of the instrument to parents of 3-month-old infants who have difficulty providing ratings for behaviors they have not yet observed.

It seems unlikely that missing data (or more specifically NA responses) can be avoided all together, especially when assessing very young infants. Fortunately, researchers now have access to a variety of new techniques for dealing with missing data that produce much better results than traditional missing data techniques when the amount of missing data exceeds 5 or 10% [1]. Our own test of two modern missing data techniques, EM and MI, suggests that these

approaches reduce many of the problems associated with NA responses. Overall, the EM approach did the best job of eliminating differences between the No NA and NA ≥ 30% groups.

The choice of a missing data strategy should be determined by the nature of the missing data. We refer interested readers to excellent and accessible reviews of modern missing data techniques by Shafer and Graham [3] and Widaman [18]. Widaman's paper, in particular, offers a principled approach to selecting an appropriate technique. Users of the IBQ-R are encouraged to examine patterns of missing data in order to determine an appropriate remedy that addresses the needs of the analyses they wish to conduct.

5.0 Conclusions

The primary goals of this paper were to highlight the fact that the standard scoring procedure for the IBQ-R is a *de facto* missing data procedure that may or may not be appropriate depending on the pattern of missing data and to examine alternate methods for handling missing data. Substituting the mean of the available items for the mean that would have been obtained if data were complete degrades the reliability and validity of multi-item measures, at least when the amount of missing data is as high as 22%. Our analysis of the short form of the IBQ-R among a community sample of infants suggests that researchers should exercise caution when interpreting results obtained from infants at 3-months of age. Careful selection of scales, selecting a full length version of the IBQ-R, and the use of modern missing data techniques may help to maintain the quality of data obtained from younger infants.

Conflict of interests

The authors declare that they do not have financial or non-financial conflict of interests.

Authors' contributions

GG and DD were involved in the conception and design of the study and the acquisition of the data. GG conducted data analysis and wrote the first draft of the manuscript. DD revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the participants of the Alberta Pregnancy Outcomes and Nutrition study and the support of the APrON Study Team, whose individual members are Bonnie J. Kaplan, Catherine J. Field, Deborah Dewey, Rhonda C. Bell, Francois P. Bernier, Marja Cantell, Linda M. Casey, Misha Eliasziw, Anna Farmer, Lisa Gagnon, Gerald F. Giesbrecht, Laki Goonewardene, David W. Johnston, Libbe Kooistra, Nicole Letourneau, Donna P. Manca, Jonathan W. Martin, Linda J. McCargar, Maeve O’Beirne, Victor J. Pop, and Nalini Singhal. This research was supported in part by grants from Alberta Innovates Health Solutions and the Alberta Children’s Hospital Foundation. The sources of funding had no role in the study design, in the collection, analysis or interpretation of data; in writing the manuscript; or in the decision to submit the manuscript for publication.

References

[1] Fichman M, Cummings J. Multiple imputation for missing data: making the most of what you know. Organizational Research Methods. 2003;6:282-308.

[2] McKnight PE, McKnight KM, Sidani S, and Figueredo AJ. Missing Data: A Gentle Introduction. New York; London: The Guilford Press; 2007.

[3] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods. 2002;7:147-77.

[4] Gartstein MA, Rothbart MK. Studying infant temperament via the Revised Infant Behavior Questionnaire. Infant Behavior & Development. 2003;26:64-86.

[5] Rothbart MK. Measurement of temperament in infancy. Child Development. 1981;52:569-78.

[6] Putnam SP, Helbig AL, Gartstein MA, Rothbart MK, Leerkes E. Development and Assessment of Short and Very Short Forms of the Infant Behavior Questionnaire-Revised. Journal of personality assessment. 2013.

[7] Kaplan BJ, Giesbrecht GF, Leung BM, Field CJ, Dewey D, Bell RC, et al. The Alberta Pregnancy Outcomes and Nutrition (APrON) cohort study: rationale and methods. Maternal & child nutrition. 2014;10:44-60.

[8] Woythaler MA, McCormick MC, Smith VC. Late preterm infants have worse 24-month neurodevelopmental outcomes than term infants. Pediatrics. 2011;127:e622-9.

[9] Acock AC. Working with missing values. Journal of Marriage and Family. 2005;67:1012-28.

[10] Downey RG, King C. Missing data in Likert ratings: a comparison of replacement methods. The Journal of General Psychology. 1998;125:175-91.

[11] Roth PL, Switzer Iii FS, Switzer DM. Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. Organizational Research Methods. 1999;2:211-32.

[12] Graham JW. Missing data analysis: making it work in the real world. Annual review of psychology. 2009;60:549-76.

[13] Gottschall AC, West SG, Enders CK. A comparison of item-level and scale-level multiple imputation for questionnaire batteries. Multivariate Behavioral Research. 2012;.47:pp.

[14] Cox JL, Holden JM, Sagovsky R. Detection of postnatal depression: development of the 10-item Edinburgh Postnatal Depression Scale. The British Journal of Psychiatry. 1987;150:782-6.

[15] Derogatis LR. Symptom Checklist-90-R: Administration, Scoring, and Procedures Manual. Minneapolis, MN: Pearson; 1994.

[16] Crocker L, Algina J. Introduction to classical and modern test theory. Philadelphia: Harcourrt Brace Jovanovich College Publishers; 1986.

[17] Graham JR, Naglieri JA. Handbook of psychology: Assessment psychology, Vol. 10. Hoboken, New Jersey: John Wiley & Sons Inc; 2003.

[18] Widaman KF. Missing data: What to do with or without them. Monographs of the Society for Research in Child Development. 2006;71:1540-5834.

Table 1. Mean (SD) number and range of NA responses for IBQ-R scales

Scale	No. of items	Number of NA responses			
		3-Months (n = 401)		6-Months (n = 325)	
		Mean (SD)	Range	Mean (SD)	Range
Activity level	7	.42 (.88)	0 - 5	0.03 (0.2)	0 – 2
Distress to limitation	7	1.0 (1.1)	0 - 5	0.3 (0.7)	0 – 4
Fear	6	1.1 (1.6)	0 – 6	0.5 (0.9)	0 – 6
Duration of Orienting	6	2.1 (1.8)[a]	0 – 6	0.4 (1.0)	0 – 6
Smiling and laughter	7	1.8 (1.4)	0 – 7	0.4 (0.6)	0 – 3
High pleasure	7	2.5 (1.9)[a]	0 – 7	0.3 (0.7)	0 – 5
Low pleasure	7	1.5 (1.3)	0 – 6	0.8 (1.0)	0 – 5
Soothability	7	0.5 (1.0)	0 – 5	0.3 (0.8)	0 – 5
Falling reactivity	6	0.3 (0.7)	0 – 5	0.1 (0.3)	0 – 2
Cuddliness	6	0.3 (0.7)	0 – 4	0.2 (0.4)	0 – 2
Perceptual sensitivity	6	3.2 (2.0)[b]	0 – 6	1.9 (1.7)[a]	0 – 6
Sadness	6	1.2 (1.4)	0 – 6	0.3 (0.7)	0 – 4
Approach	6	3.2 (1.9)[b]	0 – 6	0.7 (1.0)	0 – 6
Vocal Reactivity	7	1.1 (1.2)	0 - 7	0.5 (0.7)	0 - 6

[a] Average number of NA response was greater than 30%.

[b] Average number of NA response was greater than 50%.

Table 2a. Cronbach's alpha at T1 (3-months) calculated for different missing data techniques

Scale	No NA group (% of sample used to calculate α)	Missing Data Strategy (NA ≥ 30% group)				
		SMS (no. of items)	PMI	NA = Never	EM	MI
Activity Level	.70 (75%)	.70 (7)	.86	.23	.64	.59
Distress to Limitation	.73 **(48%)**	.70 (5)	.89	.57	.75	.71
Fear	.61 (58%)	.61 (6)	.92	.37	.71	.78
Duration of Orienting	.85 **(28%)**	.57 (2)	.94	.40	.75	.65
Smiling and Laughter	.76 **(25%)**	.61 (3)	.90	.54	.68	.64
High Pleasure	.79 **(25%)**	.30 (2)	.92	.46	.76	.67
Low Pleasure	.74 **(24%)**	.72 (5)	.88	.15	.68	.62
Soothability	.86 (72%)	.86 (7)	.89	.52	.78	.75
Falling reactivity	.82 (79%)	.82 (6)	-	-	-	-
Cuddliness	.74 (73%)	.74 (6)	-	-	-	-
Perceptual sensitivity	.85 **(15%)**	- (0)	.96	.48	.71	.57
Sadness	.75 **(44%)**	.64 (4)	.91	.35	.72	.60

Approach	.87 **(15%)**	.60 (2)	.96	.39	.65	.52
Vocal Reactivity	.72 **(37%)**	.69 (6)	.83	.39	.70	.59

Note: No NA group = infants with no responses marked NA; NA ≥ 30% group = infants with 2 or more items on a scale marked as NA. Bolded cells indicate that alphas were based on less than 50% of the sample. Blank cells indicate that either the No NA group or the NA > 30% group had a samples size less than 10% of all cases. Data in the SMS column are from scales that eliminated *items* with high rates of NA responses (≥ 30% of the sample); No. of items in the SMS column refers to the number of remaining items after deleting items with 30% or more NA responses.

Table 2b. Cronbach's alpha at T2 (6-months) calculated for different missing data techniques

Scale	No NA (Percentage of sample used to calculate α)	Missing Data Strategy (NA responses ≥ 30%)				
		SMS (no. of items)	PMI	NA = Never	EM	MI
Activity Level	.71 (94%)	.71 (7)	-	-	-	
Distress to Limitation	.74 (76%)	.74 (7)	.91	.76	.80	.75
Fear	.78 (66%)	.78 (6)	-	-	-	
Duration of Orienting	.78 (73%)	.78 (6)	.94	.46	.88	.75
Smiling and Laughter	.77 (61%)	.74 (6)	-	-	-	
High Pleasure	.73 (78%)	.73 (7)	-	-	-	
Low Pleasure	.70 (51%)	.70 (7)	.86	-.16	.72	.65
Soothability	.84 (83%)	.84 (7)	-	-	-	
Falling reactivity	.81 (88%)	.81 (6)	-	-	-	
Cuddliness	.82 (80%)	.82 (6)	-	-	-	
Perceptual sensitivity	.84 **(27%)**	.52 (2)	.93	.40	.75	.63
Sadness	.68 (72%)	.68 (6)	-	-	-	

Approach	.68 (57%)	.68 (6)	.91	.27	.77	.68
Vocal Reactivity	.79 (51%)	.71 (6)	-	-	-	

Note: No NA group = infants with no responses marked NA; NA ≥ 30% group = infants with 2 or more items on a scale marked as NA. Bolded cells indicate that alphas were based on less than 50% of the sample. Blank cells indicate that either the No NA group or the NA > 30% group had a samples size less than 10% of all cases. Data in the SMS column are from scales that eliminated *items* with high rates of NA responses (≥ 30% of the sample); No. of items in the SMS column refers to the number of remaining items after deleting items with 30% or more NA responses.

Table 3a. Means (standard deviations) at T1 (3-months) calculated for different missing data techniques.

	No NA group	Missing Data Strategy (NA ≥ 30% group)			
		SMS & PMI	NA = Never	EM	MI
Activity Level	3.29 (0.96)	3.31 (1.19)	2.44 (.74)***	3.21 (.99)	3.31 (.16)
Distress to Limitation	3.72 (0.99)	3.86 (1.17)	3.10 (.85)***	3.88 (1.09)	3.84 (.10)
Fear	2.00 (0.62)	2.90 (1.27)***	1.70 (.48)***	1.97 (1.23)	2.11 (.11)
Duration of Orienting	3.76 (1.47)	4.15 (1.38)*	2.28 (.84)***	3.79 (1.23)	3.78 (.09)
Smiling and Laughter	3.68 (1.23)	3.95 (1.33)†	2.70 (.85)***	3.69 (1.16)	3.68 (.09)
High Pleasure	4.59 (1.26)	4.80 (1.28)	2.88 (.99)***	4.47 (1.20)	4.43 (.07)
Low Pleasure	5.34 (0.97)	5.62 (0.93)*	3.76 (.82)***	5.33 (.87)	5.33 (.07)
Soothability	5.45 (0.94)	5.49 (1.00)	5.20 (.62)**	5.43 (.86)	5.39 (.10)
Falling Reactivity	5.14 (1.04)	4.98 (1.40)	-	-	-
Cuddliness	6.18 (0.64)	6.06 (0.91)	-	-	-
Perceptual Sensitivity	2.74 (1.52)	3.39 (1.80)**	1.70 (.77)***	3.03 (1.42)	3.18 (.09)*
Sadness	3.40 (1.13)	3.53 (1.32)	2.31 (.79)***	3.23 (1.21)	3.27 (.10)
Approach	3.63 (1.44)	3.64 (1.71)	1.81 (.77)***	3.44 (1.21)	3.52 (.07)
Vocal Reactivity	4.14 (1.05)	3.90 (1.09)†	2.76 (.85)***	3.65 (1.09)***	3.76 (.10)**

Note: No NA group = infants with no responses marked NA; NA ≥ 30% group = infants with 2 or more items on a scale marked as NA. Mean comparisons were conducted using independent samples T-tests; the No NA group is the comparison group for all analyses. *** < .001; ** < .01; * < .05; † < .10. For the MI column, values in brackets are standard error. Blank cells indicate that either the No NA group or the NA > 30% group had a samples size less than 10% of all cases.

Table 3b. Means (standard deviations) at T2 (6-months) calculated for different missing data techniques.

	No NA	Missing data strategy for infants with high NA responses (≥ 30%)			
		SMS & PMI	NA = Never	EM	MI
Activity Level	3.98 (0.98)	-	-	-	-
Distress to Limitation	3.45 (0.95)	3.41 (1.17)	2.86 (.99)***	3.32 (1.07)	3.41 (.20)
Fear	2.40 (1.05)	-	-	-	-
Duration of Orienting	4.20 (1.15)	4.19 (1.31)	2.71 (.99)***	3.99 (1.38)	4.19 (.22)
Smiling and Laughter	4.41 (1.09)	-	-	-	-
High Pleasure	5.97 (0.73)	-	-	-	-
Low Pleasure	5.52 (0.82)	5.52 (0.90)	3.95 (.74)***	5.41 (.87)	5.51 (.10)
Soothability	5.74 (0.87)	-	-	-	-
Falling Reactivity	5.38 (1.00)	-	-	-	-
Cuddliness	5.88 (0.84)	-	-	-	-
Perceptual Sensitivity	3.73 (1.45)	4.19 (1.50)*	2.44 (.91)***	3.79 (1.41)	4.19 (.12)*
Sadness	3.17 (0.96)	-	-	-	-
Approach	5.15 (0.86)	4.61 (1.29)***	3.10 (.97)***	4.68 (1.15)***	4.61 (.18)***
Vocal Reactivity	4.90 (1.02)	-	-	-	-

Note: No NA group = infants with no responses marked NA; NA ≥ 30% group = infants with 2 or more items on a scale marked as NA. Mean comparisons were conducted using independent samples T-tests; the No NA group is the comparison group for all analyses. *** < .001; ** < .01; * < .05; † < .10. For the MI column, values in brackets are standard error. Blank cells indicate that either the No NA group or the NA ≥ 30% group had a samples size less than 10% of all cases.

Table 4. Longitudinal stability (Pearson correlation) observed between 3- and 6-months for infants with complete data and using different missing data techniques

	No NA group		Missing Data Strategy (NA ≥ 30% group)				
Scale	r	n	n	SMS & PMI	NA = Never	EM	MI
Activity Level	.42***	247	31	.41*	.31	.49***	.34
Distress to Limitation	.47***	160	120	.46***	.48***	.41***	.42***
Fear	.43***	195	84	.16	.20	.32***	.14
Duration of Orienting	.56***	95	186	.31***	.32 ***	.36***	.23*
Smiling and Laughter	.50***	79	184	.48***	.46***	.46***	.49***
High Pleasure	.51***	85	227	.32***	.31***	.47***	.35***
Low Pleasure	.70***	81	148	.37***	.39***	.40***	.31***
Soothability	.27***	239	65	.59***	.43***	.84***	.56***
Falling reactivity	.34***	260	26	.06	-.05	.12	.06
Cuddliness	.20**	244	20	.29	.15	.36	.33
Perceptual sensitivity	.49***	51	232	.45***	.44***	.46***	.33***
Sadness	.47***	145	105	.39***	.33***	.46***	.35**
Approach	.49***	48	255	.25***	.23***	.32***	.21*

Vocal Reactivity	.59***	122	88	.23*	.28**	.38***	.25*
Mean across all scales	.46			.34	.31	.42	.31

Note: No NA group = infants with no responses marked NA; NA ≥ 30% group = infants with 2 or more items on a scale marked as NA. Asterisk indicates that the test-retest reliability is statistically significant *** < .001; ** < .01; * < .05; † < .10.