|  |  |  |  |
| --- | --- | --- | --- |
| **Supplemental Table 1:** Definition and Interpretation of Key Statistics Produced During Rasch Analysis | | | |
| ‘Indicator’ (Statistic) | Definition of indicator and rationale for its assessment | Statistic in RUMM | Interpretation of statistic in RUMM |
| Overall fit to the Rasch Model | In a Rasch model, it is assumed that responses to individual items will depend on the underlying level of knowledge that individual has. The actual fit of (a) persons and (b) items, to these assumptions must be assessed statistically, to determine whether modification of individual items is required, or individual responses need to be deleted from the analysis | Fit-residual statistic (sum of all differences between observed and expected values)  Chi Square probability | Z score of 1 and SD of 0 = perfect fit; SD > 1.5 = misfit and further assessment required  A significant P-value (< 0.05 or Bonferoni adjusted value) for chi-square probability = misfit and further assessment required |
| Person Separation Index (PerSepIndex) | PerSepIndex gives an indication of internal reliability based on the basic definition given in CTT | Person separation index (PSI) | PSI > 0.7 = acceptable (is interpreted in a similar manner to Cα ) |
| Local dependency | Local dependency occurs when the answer to one item is ‘dependent’ on the answer to another item *(e.g. a distractor option in a multi-choice question may inadvertently give away the answer to a true/false question)* | Residual correlations | R > 0.2 = items need to be reviewed; locally dependent items can be formed together as their own nutrition sub-section(s) with overall fit statistics re-run – if these are similar to the original results it can be concluded that local dependency is not a problem |
| Unidimensionaility (produces similar results to factor analysis) | Unidimensionality assesses that all items are assessing the same ‘construct’ *e.g. that nutrition knowledge can be thought of as a single construct, meaning that the individual questions are all assumed to equally understood and knowledge of a particular nutrition topics being assessed by the questionnaire, such as hydration, is assumed to be related (and equal) to knowledge of another topic being assessed in the questionnaire, such as supplementation* | Principal component analysis (PCA) on residuals (PC loadings)  T-test and PerC <5%; If two ‘sets’ (which can be thought of as nutrition sub-sections) are formed after PCA, the items in each set can be selected and a t-test can be run for each person in the sample to compare scores on ‘set 1’ and ‘set 2’ | PC loading > 0.2 within each principal component indicate a set or ‘nutrition sub-section’ is present  A significant P-Value (<0.05) indicates scores on the two sets are statistically different and that the questionnaire may be multidimensional; if more than 5% of persons get a P-value of <0.05 i.e. if the ‘Perc<5%’ statistic is greater than 5.0% than the questionnaire may need to be separated into multiple sets, and Rasch analysis performed on these sets. |
| Differential item functioning (DIF); also referred to as ‘Item Bias’ | DIF occurs when an item is answered differently (by two individuals who actually have the same level of knowledge) because of differences in characteristics such as age and gender *(e.g. a male may answer a question that refers to cars better than female with the same amount of knowledge)*. DIF can be uniform (i.e. systematic, that is males always respond differently to females) or non-uniform (i.e. random) | Item characteristic curve (plots person location against ‘expected value’)  DIF summary statistics (DF, P-value) | Visual interpretation of differences based on person characteristics  A significant P-value (< 0.05 or Bonferroni adjusted value) = significant DIF  If uniform DIF is found, it may be necessary to change the scoring of items based on the characteristic causing the DIF; if non-uniform DIF is found the item may need to be deleted |
| Response thresholds (if Likert Scales or multi-choice items have been used) | ‘Threshold’ refers to the point between two responses options were choosing either option is equally likely. Assessment of thresholds is performed to ensure that respondents are consistently able distinguish between response categories (e.g. very likely, likely). *This is mostly relevant to attitude style tests and is unlikely to have utility in achievement tests such as measures of nutrition knowledge* | (i) Threshold map  (ii) Category probability curves (plots ‘person location’ i.e. range in the trait being measured, against probability of endorsing a particular response)  (iii) Category frequencies (charts frequency with which each response category has been selected) | Map missing from output = disordered threshold and investigate further with item (ii) and (iii)  If a particular response category does not appear have the highest probability of being endorsed at least once = disordered thresholds  If a particular category is not being ‘selected’ enough it may need to be removed e.g. combine ‘likely and very likely’ and make a 4-point rather than 5-point Likert scale |