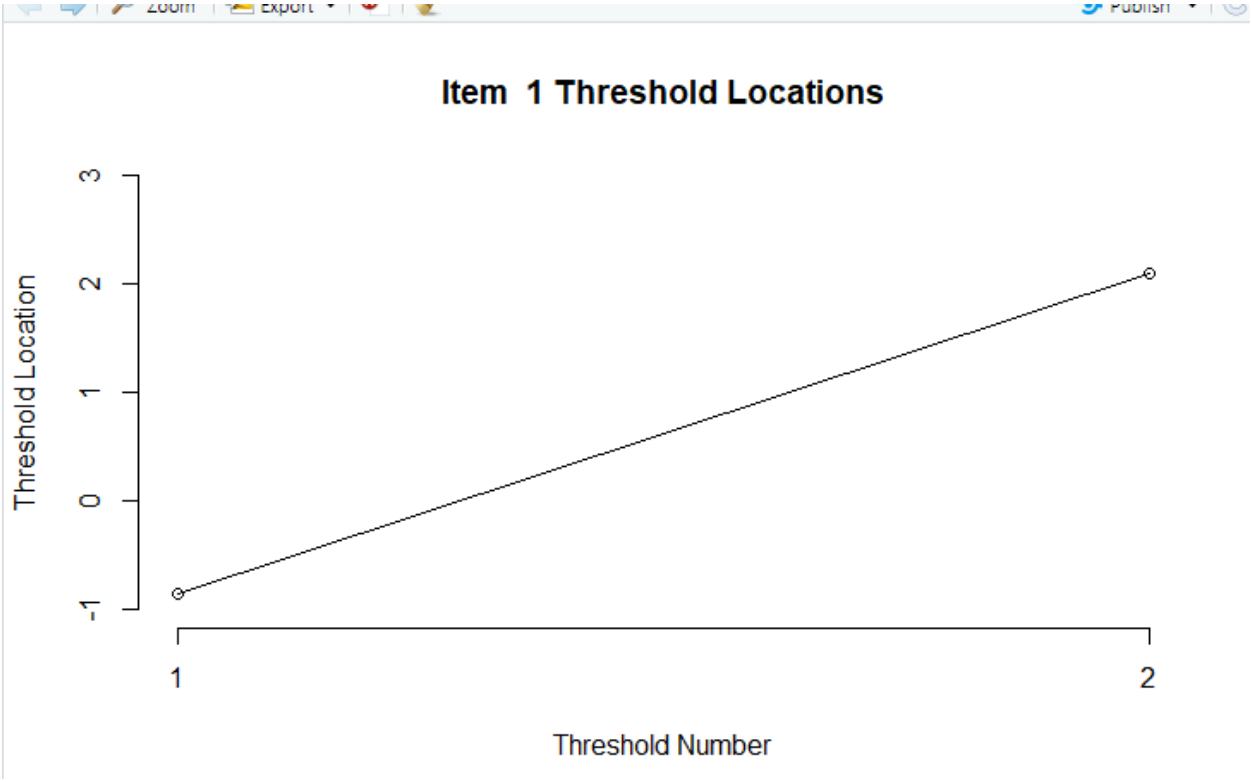
There are loads of other analysis which can be performed using the winsteps software. They are not mentioned here

First we would examine the data to see if there are any missing values or other potentially problematic things.

```
 Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.
> apply(survey.responses,2,table)
  LeadPetrol RiverSea RadioWaste AirPollution Chemicals Nuclear
0         17        7         18           9        17      46
1         95       51         56          93        56      95
2        179      233        217         189       218     150
>
>
```

This shows the frequency of responses in each category. We might look for any categories with too few responses. A recommended minimum number of responses per category is 10 (Linacre, 1999).

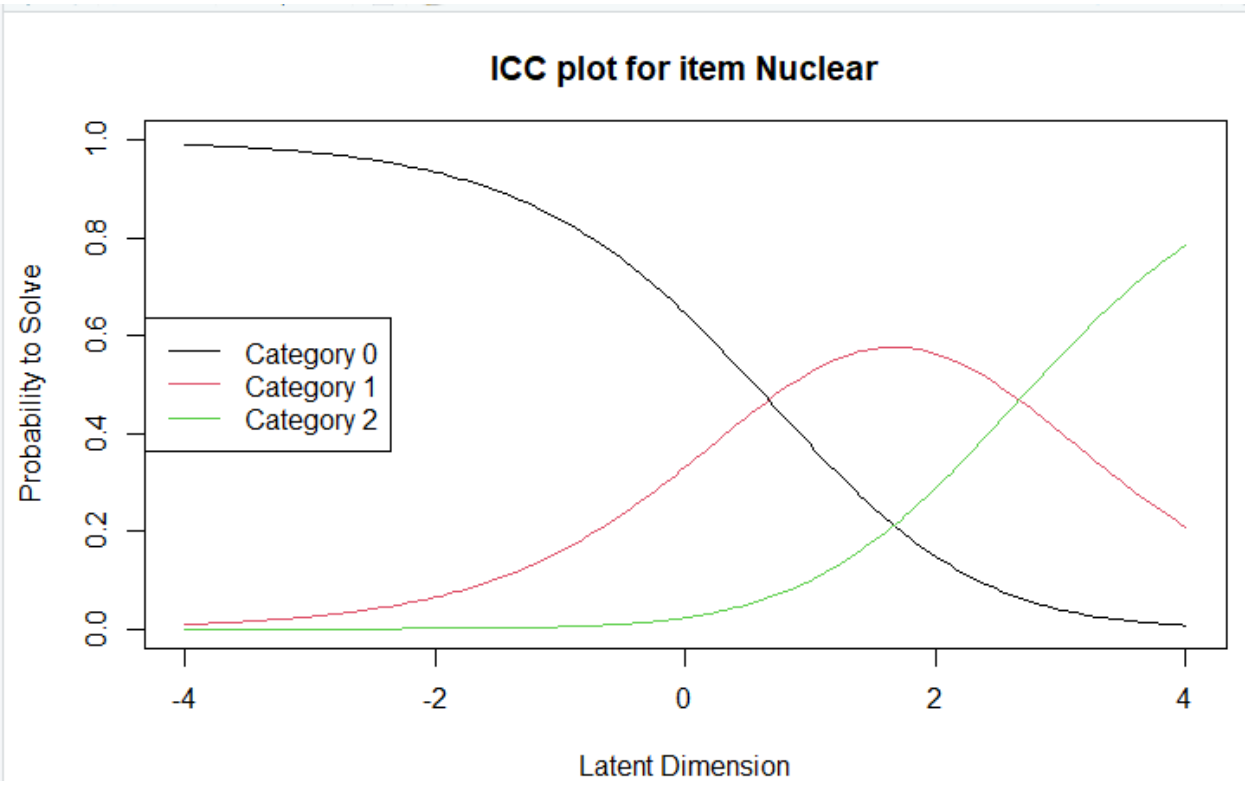


We can Check category threshold estimates for non-decreasing (i.e., monotonic) ordering over increasing categories for all items. As for average measures, step calibrations should increase monotonically. Thresholds that do not increase monotonically across the rating scale are considered disordered.

We can check the following plots for

- Ordered category curves
- Distinct (modal) categories

Each response category should have a distinct peak in the probability curve graph, illustrating that each is indeed the most probable response category for some portion of the measured variable. Categories observed to be ‘flat’ on the graph might be useful as long as they span a large portion of the variable. If, however, these flat categories are overshadowed and made redundant by other categories, they might not aid in defining a distinct point on the variable. Problematic thresholds—those that are disordered or too close—will show up on the graph, often as flattened probability curves spanning small sections of the measured variable.



Other analysis
Two features are important in the category frequencies: shape of the distribution and number of responses per category. Regular distributions such as uniform, normal, bimodal, slightly skewed distributions are preferable to those that are irregular. Irregular distributions include those that are highly skewed. If any categories have low responses we might considering collapsing those categories

Average measures are useful for 'eyeballing' initial problems with rating scale categories. Average
measures are defined as the average of the ability estimates for all persons in the sample who chose that particular response category, with the average calculated across all observations in that category (Linacre, 1995). For example, if the average measure is −1.03 logits for Category 1, that is the average of the ability estimates for all persons who chose Category 1 on any item in the survey. Average measures should increase as the variable increases. If average measures increase monotonically across response categories, that indicates that, on average, those with higher ability/stronger attitudes endorse progressively higher categories, whereas those with lower abilities/weaker attitudes endorse progressively lower categories. When this pattern is violated, as indicated by a lack of monotonicity in the average measures, collapsing categories might be considered.
The threshold estimates should be neither too close together nor too far apart on the logit scale. Guidelines recommend that thresholds should increase by at least 1.4 logits to show empirical distinction between categories but not more than 5 logits so as to avoid large gaps in the variable

Fit statistics provide another criterion for assessing the quality of rating scales. Outfit mean squares
greater than 2 indicate more misinformation than information (Linacre, 1999), meaning that the particular category is introducing more noise than meaning into the measurement process. Such categories warrant further empirical investigation and thus might be considered as candidates for collapsing with adjacent categories. Table 11.3 shows the fit of each rating scale category to the unidimensional Rasch model, well under the criterion of mean square statistic less than 2.0

IBut what is the evidence for deciding which is the better of the two when we have to collapse up and down? When comparing several categorizations of the same rating scale, we also can look at indicators
other than category diagnostics. For example, we can assess the quality of the various reliability and validity indices for the whole variable and compare these across each recategorization. With respect to validity, we will look at both the item ordering and fit. That is, does one categorization of the variable result in a better ordering of items along the underlying variable, one that is more consistent with the theory that generated the items in the first place? Do items misfit under one categorization but not another? These reliability and validity issues are addressed in previous chapters, but the same principles continue to apply here as well. Here they help us gain a fuller picture of how we can best refine the rating scale. Whereas the rating

scale diagnostics help us in determining the best categorization, knowledge of Rasch reliability and validity indices tells us how the measure is functioning as a whole.