Memo: UNC STTR Paper Update
Date: 2.15.17

FROM: Andrea Hussong

Purpose of project: Examine select sources of cross-study variation in a harmonized measure of drug use and consequences based on 11 STTR studies (from the Criminal Justice System subset) using three methods of analytic harmonization to create commensurate scores.

Proposal Title: Alternative approaches for analytic harmonization of drug use symptoms from 11 STTR studies

Chair Person: XXXX

All author names: XXXX

Aims/ Hypotheses

In the proposed paper, we contrast four scoring approaches for creating commensurate measures across the 8 CJ STTR sites (in which we collapsed studies from the same site that used the same instrumentation – these are identified by STTR documentation, see below). These approaches include scores based on:

(1) a moderated non-linear factor analysis (MNLFA),
(2) a traditional confirmatory factor analysis (that does not take into account differential item functioning, traditional CFA),
(3) sum scores from logically harmonized AUDIT items
(4) AUDIT-based cut-offs for problem drinking.

In this manuscript, we will evaluate the magnitude and practical importance of the differences between scores generated using different methods with a complex dataset. Take home implications of this work include answers to the following questions about STTR: (a) can the MNLFA approach to creating commensurate measures across the CJ STTR data sets be used to retain all available items across studies even when those items were administered in different ways or are only available in some of the data sets, (b) how do scores obtained from the MNLFA approach compare to scores obtained from more traditional approaches that are more limited in their ability to address cross study differences in item administration, and (c) how do the items on the alcohol consequences measure behave differently (i.e., are differentially related to the underlying alcohol consequences factor) based on study membership as well as gender, race/ethnicity, and HIV status of participants. In addition, we will examine (d) differences in predictive validity of the four types of scores.

Sample Descriptions

We included data from a total of 8 CJ STTR study sites. We excluded the 3 studies because the AUDIT was not used to assess alcohol problems and 1 study because AUDIT responses were not available in the original data set shared with us. The remaining studies are listed in Table 1 along with the sample size and demographic characteristics as found in the data set we received.

Table 1. Demographics by Study PI / Sample

Study PI	Sample	*N*	% Male	%Black	% Asian	% White	% Multiethnic	Age M (SD)
Gordon	Bright1	2405	82	52	0	28	4	38.99 (29.96)
	Bright2	100	78	90	0	3	3	45.22 (7.76)
Beckwith	Careplus-parole	138	80	19	0	41	6	34.41 (10.30)
	Careplus-rct	112	58	88	0	4	8	32.90 (10.16)
	Careplus-ridoc	250	94	15	0	47	8	42.23 (9.56)
Springer	Newhope	122	81	23	0	12	1	45.64 (8.09)
Sacks	Starttogether	195	100	46	1	22	1	37.90 (10.86)
Altice	Stride1	50	69	94	0	0	4	53.02 (6.36)
	Stride2	109	46	37	0	1	0	50.83 (9.04)
Ouellet	Sttcoip-jail	376	80	81	0	8	1	40.91 (12.02)
	Sttcoip-prison	90	88	91	0	3	0	43.72 (10.34)
Spaulding	Success	56	89	70	0	9	7	37.16 (8.61)
Quan	Vista	378	100	0	100	0	0	35.58 (5.97)

Findings from Descriptive Item and Factor Analyses

We examined item distributions within and across studies and conducted a series of exploratory factor analyses to assess dimensionality and local dependence (redundant items). Based on these analyses, we selected items and transformed variables as follows in preparation for scoring.

- To create MNLFA, CFA, and sum scores, we transformed all items to a binary response scale (yes/no) to avoid sparseness. We also eliminated quantity and frequency items in MNLFA, CFA and sum scores because they were redundant with the binge item and because they split off to form a separate factor (with the binge item) in an EFA. Without these items, the remaining items form a unidimensional factor and were used in sum score and MNLFA scoring: Cutdown, Injury, Forget, Guilt, Morning, Fail normal, Can't stop, and Binge. Percent endorsement of AUDIT items are listed for each study in Table 2.

Table 2. Item endorsements by Study PI

	21 (Care)	25 (NewHope)	26 (STT)	27 (Bright)	28 (Start)	29 (Stride)	31 (Success)	33 (Vista)
BINGE	61%	28%	49%	29%	54%	31%	48%	53%
NOSTOP	37%	20%	24%	10%	32%	18%	27%	100%
FAILNORM	33%	20%	25%	11%	27%	20%	30%	24%
MORNING	27%	19%	15%	7%	24%	17%	21%	9%
GUILT	36%	26%	25%	15%	32%	19%	30%	22%
FORGET	37%	18%	20%	15%	24%	19%	30%	10%
INJURY	29%	6%	18%	12%	20%	8%	11%	21%
CUTDOWN	37%	21%	26%	19%	33%	8%	25%	60%

- To create AUDIT-based cutoff scores, we shifted response scales from all studies to align with AUDIT response scoring (0-4 scale for items 1-8 and responses of 0, 2, or 4 for items 9-10) and then summed across all ten items. The standard cut-off of score of 8 or more vs. less than 8 was used to create a dichotomous indicator (0/1) of hazardous/harmful alcohol use.
    - Note: For Bright1/Bright2 studies, items 9 and 10 from the AUDIT had a different response scale that was not easily harmonizable with the standard AUDIT. Instead of the response options being (0) No; (2) Yes, but not in the last year; and (4) Yes, during the last year, responses were (0) Never, (1) Less than monthly, (2) Monthly, (3) Weekly, and (4) Daily or almost daily. These responses were not directly harmonizable with the standard AUDIT scale. To broadly harmonize items with AUDIT scoring, respondents who answered '1' (less than monthly) were rescored as 2 ('yes, but not in the past year') on the standard AUDIT and if they answered > 1 (more than monthly) they were scored as '4' (yes, in the past year). However, we recognize that a response of "less than monthly" could possibly indicate that it happened in the past year, making the scoring not entirely accurate for this population. (Documentation for Bright 1 & 2 suggest that an "alcohol use addendum" was also given with the standard AUDIT response scale; however, we do not appear to have these items in our datasets.)
    - All other studies had response scales harmonizable with standard AUDIT scoring. However, there were differences in studies regarding how items were asked, including the retrospection period for the AUDIT questions and whether or not binge drinking criteria was asked differently based on the biological sex of the participant (i.e. lower for women than men). Therefore, though standard AUDIT scores could be generated easily for all studies aside from Bright 1 and Bright 2, they may reflect different criteria across

studies. These differences were the reason for exploring MNLFA scoring in this manuscript.

MNLFA Results

Model details

- Along with study membership, we attempted to control for demographic factors (e.g., race/ethnicity, gender) and study characteristics (such as whether participants were asked to retrospect on drinking prior to incarceration), as these may inform item performance and enhance score precision. However, these models did not estimate, probably due to substantial inbalances across studies giving rise to excessive multicollinearity with study membership. Therefore, the MNLFA models only control for study effects.
- Study effects were effect coded. The largest study (Bright1 & Bright 2) was used as a the reference group. Study membership was allowed to have mean and variance impact on the latent factor (assessing alcohl-related consequences) as well as intercept and loading DIF (assessing differences in item performance across studies).
- Factor mean and variance impact models were run sequentially DIF was tested separately for each item. A cumulative model was then tested in which significant effects (for factor means and variances and DIF) were all included. Non-significant effects in the cumulative model were trimmed sequentially (following guidelines below) to then obtain the final model used for creating MNLFA scores.
- Trimming guidelines included the following. Any impact with $p>.10$, and intercept DIF with $p>.05$, and any intercept or loading corresponding to loading DIF with $p>.05$ was retained for a cumulative model. Any impact with $p>.10$ and any intercept or loading DIF with the loading DIF passing the 5% family-wise error false detection rate, and any additional intercept DIF passing the 5% family-wise error false detection rate, was retained in the final scoring model.

Model Results for Factor Mean (Impact) and Variance Differences by Study

- CARE, and START had positive mean impact indicating that these two studies had higher MNLFA factor scores on the AUDIT than the average score across studies. NEW HOPE had negative mean impact. Mean impact of STRIDE was included in the final scoring model (as per trimming guidelines) but was not significant in the final model
- No study had significant variance impact on the latent factor (although two were included in the final scoring model: NEW HOPE and STT).

Model Results for Item Factor Mean (Intercept) and Loading Differences by Study

- There was no significant loading DIF in the final scoring model. This indicates that across this set of studies, all items were related to the underlying factor in the same way across studies.
- STT was negatively associated with the intercept of Injury, indicating that at the same level of alcohol related consequences participants in the STT study were more likely to endorse this item than were participants in other studies (i.e., the item was easier to endorse in STT).
- VISTA was not significantly related to Fail Normal but was included in the final scoring model
- CARE and STT were positively associated the intercept for with Binge, indicating that at the same level of alcohol related consequences participants in these studies were less likely to endorse this item than were participants in other studies (i.e., the item was harder to endorse).
- NEW HOPE and STRIDE were negatively associated with Binge (easier to endorse).

Comparison of MNLFA, CFA, Sum, and AUDIT cut-off Scores

- Sum scores and CFA scores are more zero-inflated than MNLFA score have the most variance
- Figure 1 shows MNLFA (top) vs. CFA (middle) vs. sum scores (bottom) by study membership.
- The same general pattern holds for study effects across MNLFA and sum scores, though floor effects are clearly visible in sum scores and CFA scores, and to a lesser extent in MNLFA score.
    - Sum scores have a mean of 1.69 and sd = 2.41; N=310 people have missing scores
    - CFA scores have a mean of .00 and sd=.89 with no missing data due to the use of FIML to estimate scores from available items and study membership indicators.
    - MNLFA scores have a mean of -.18 and sd = .93; no scores are missing due to the use of FIML to estimate scores from available items and study membership indicators.
- In spite of the increased variability in MNLFA scores versus CFA scores, they are perfectly correlated. The correlation between sum scores and MNLFA scores is r = .90. Figure 2 shows the associations amongst the three scoring methods with univariate distributions on the diagonal.
- As shown in Figure 3, people who do not meet the cutoff criteria tend to have very low AUDIT sum scores (mean=median=mode = 0) and those who do meet criteria have a very wide range of sum scores. CFA and MNLFA scores appear to be more discriminating with respect to who meets criteria for alcohol problems.
- Figure 4 shows more detail about the pattern of MNLFA scores for people who do and do not meet the AUDIT threshold.

Summary and Next Steps

- We found that we were able to successfully create MNLFA scores for the 11 CJ STTR datasets that take into account gross study differences in item performance but we were not able to estimate more complex models that also controlled for differences in item performance due to demographic, HIV, and other factors. This was primarily due to model complexity relative to the multivariate distribution of these variables. One point of discussion is whether there is a more parsimonious set of factors we may want to consider within these MNLFA models.
- Although MNLFA and sum score were highly correlated, there were differences in score distributions - MNLFA scores had greater variability and less missing data that sum scores. though MNLFA and CFA scores were identical in rank they differ slightly in mean (due to the lack of item loading DIF in this particular well-developed scale but presence of factor impact differences). MNLFA scores also capture greater sub-threshold differences in alcohol-related consequences that fall below the AUDIT cut-score as compared to sum scores. We have found differences in variability in MNLFA and CFA scores in the past when greater item loading DIF is present, so this may not be the case for other STTR scales.
- We anticipate that the greater variability in MNLFA scores will translate to more predictive variability relative to sum scores and AUDIT cut-off scores. The next step is to examine differential predictive validity of these four types of score and whether study differences on predictive associations varies across the four types of scores. We would like to consult with co-authors on the types of validity measures that may work best in the data set and are particularly interested in those that are already logically harmonized, single item but high impact measures, or identically administered over studies.

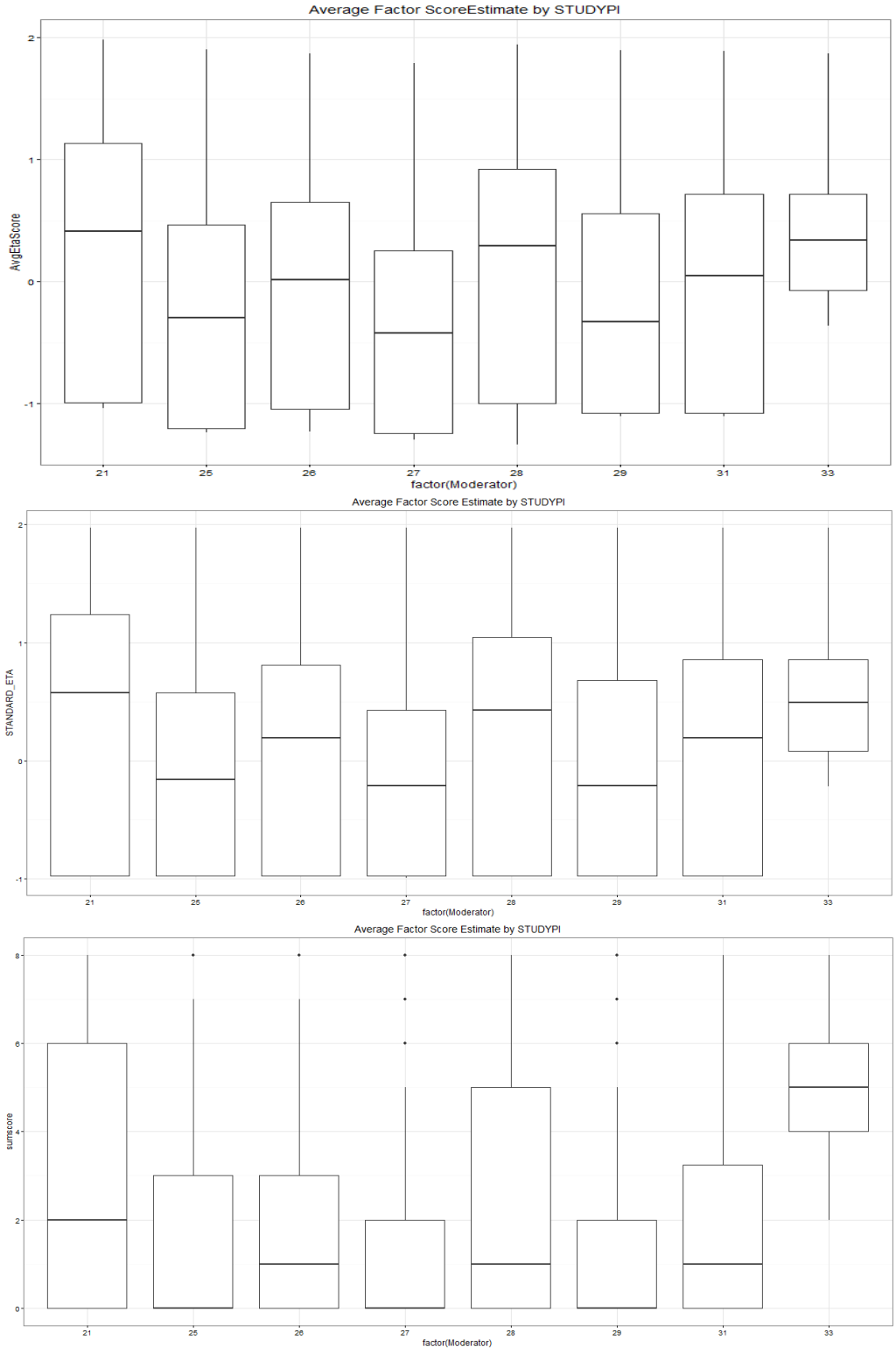

Figure 1. MNLFA scores (top), CFA scores (middle), and sum scores (bottom) plotted by study. Study membership indicated by 21=Careplus; 25=Newhope; 26=STT; 27=Bright; 28=Starttogether; 29=Stride; 31=Success; and 33=Vista.

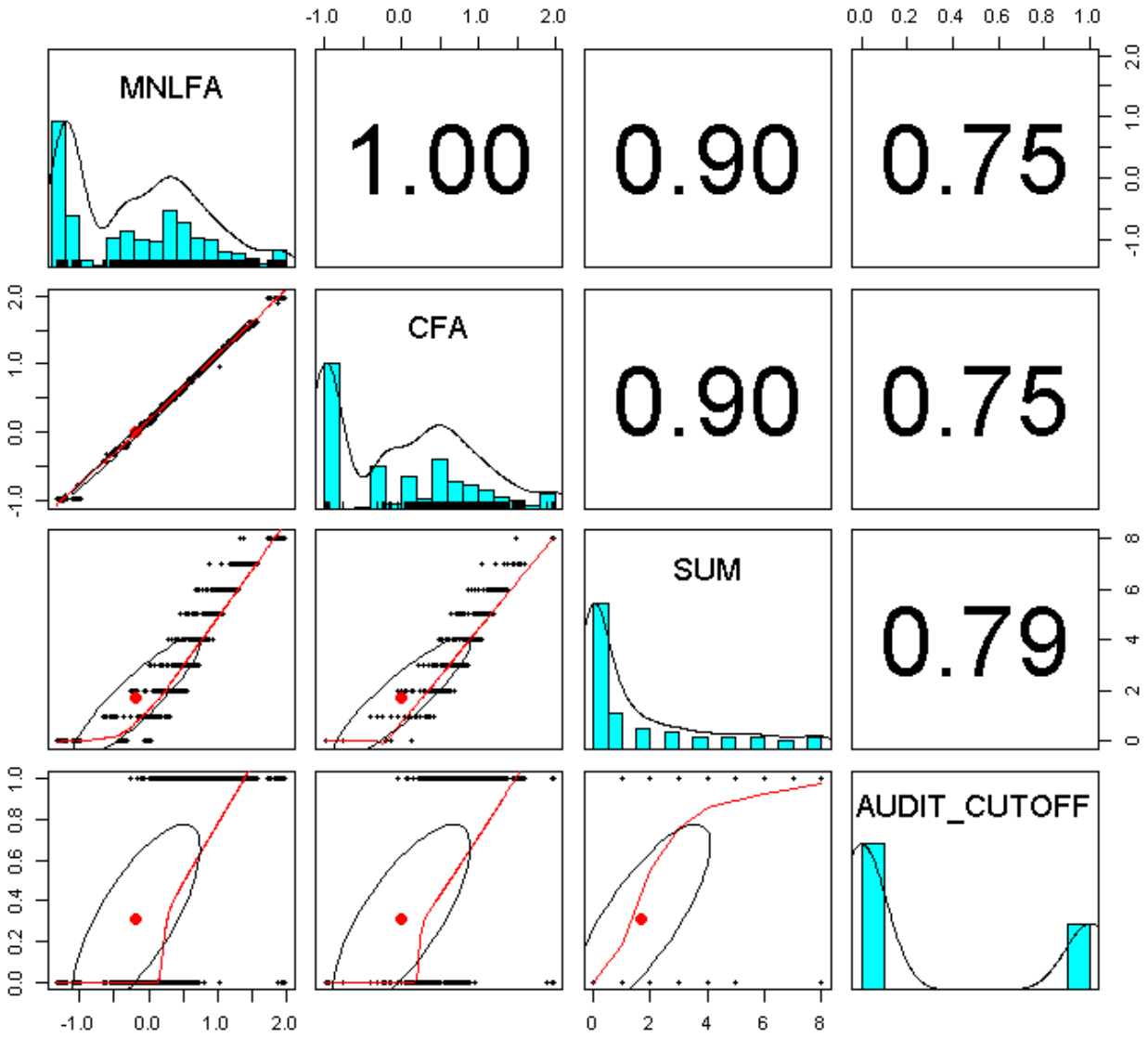

Figure 2. Comparison of MNLFA, (standard) CFA, sum scores, and AUDIT cut scores. The diagonal is the univariate distribution for each score. The upper triangle includes correlations among the scores. The lower triangle depicts bivariate associations among scores, along with a less curve and correlation ellipses to indicate the bivariate density of observations.

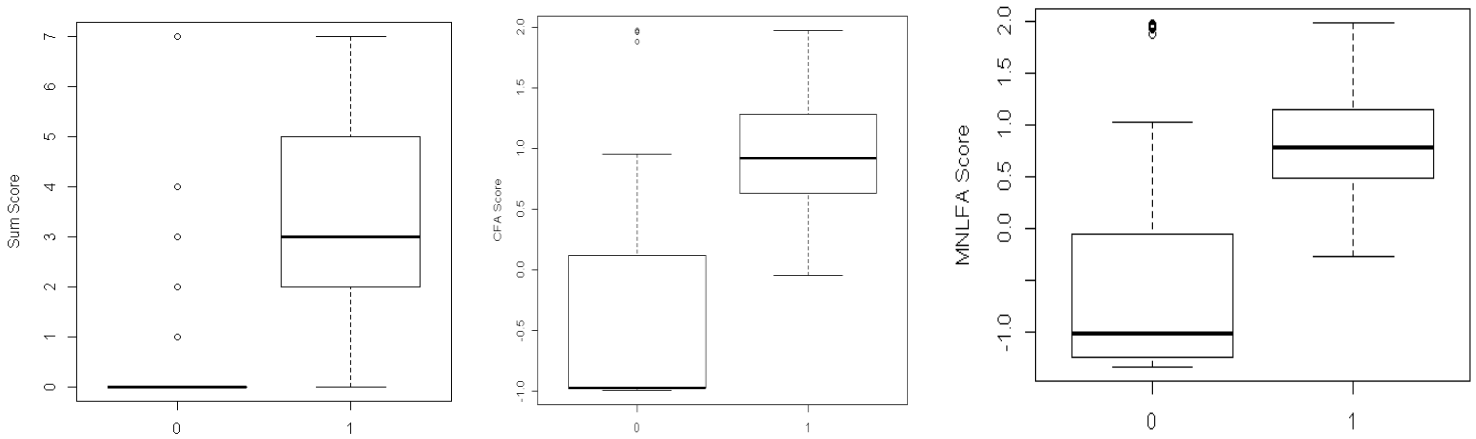

Figure 3. Distribution of sum scores (left), CFA scores (middle), and MNLFA scores (right) for those who do and do not meet AUDIT cut-score threshold

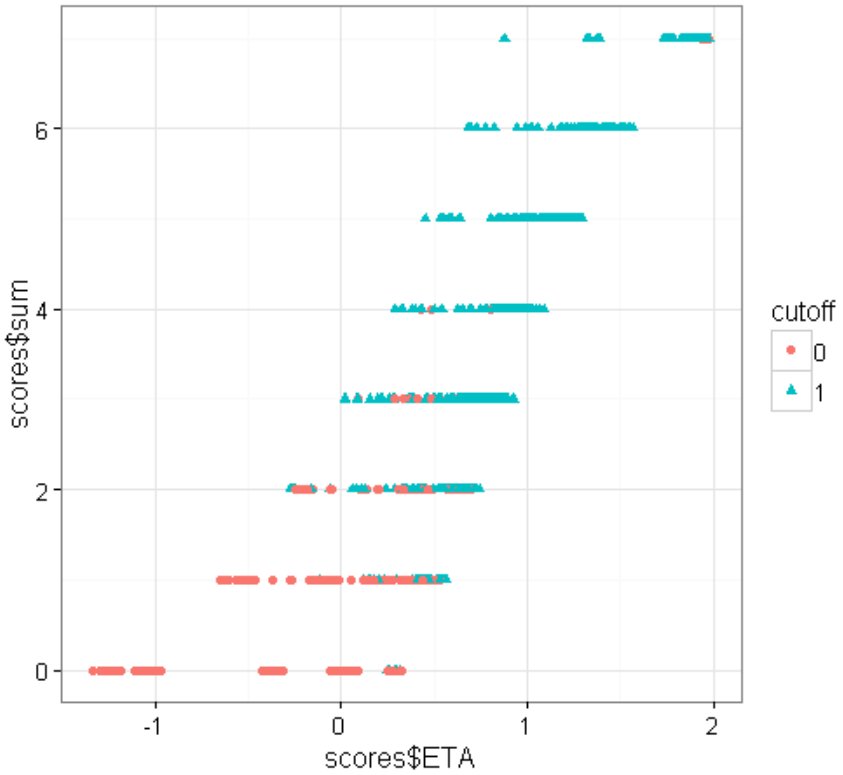

Figure 4. MNLFA scores (x-axis), sum scores (y-axis), and AUDIT cut-score (color coded, orange = below and blue=above the threshold)