Biostatistics 658: Statistics for Psychosocial Research II Problem Set #2

This problem uses the data set pset2.sav, which is in SPSS format and consistent for use with AMOS. A version of the data in Stata format is also available on the course website and is necessary for this problem set. The variable list is attached.

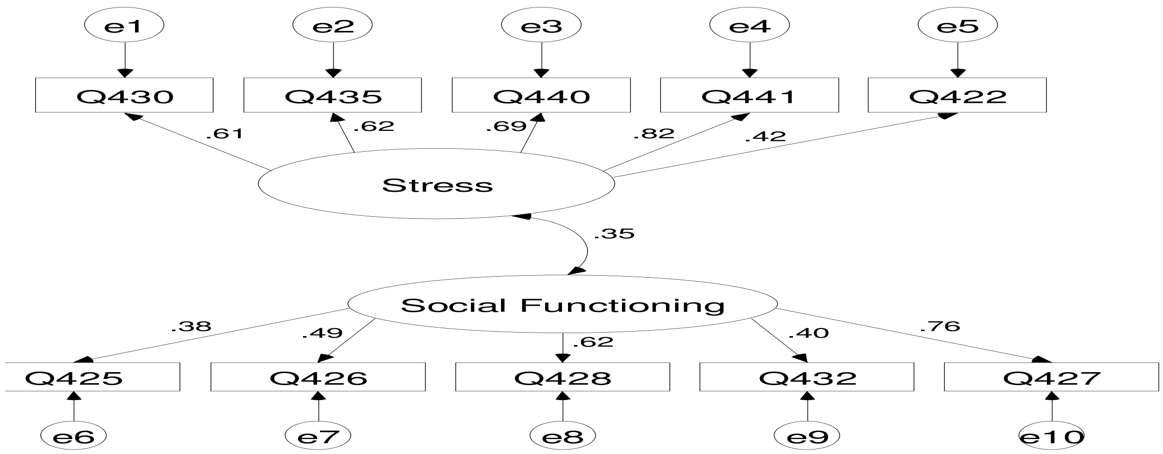
The general idea for this problem set is to estimate the effects of socioeconomic status and social mobility on mental distress. The data set is a baseline assessment of residents of East Baltimore in 1982 (the sample was later followed-up and as been funded to be reassessed in 2004).

1. Begin with the measurement aspects of mental distress. For this exercise we will treat all distress variables as continuous. First, we will choose which items should be used to measure distress.
(A) Take a subset of about half of the 20 GHQ (q422-q441) items which you predict will split into two factors (i.e. you should choose approximately 10 items). Give names to the factors.

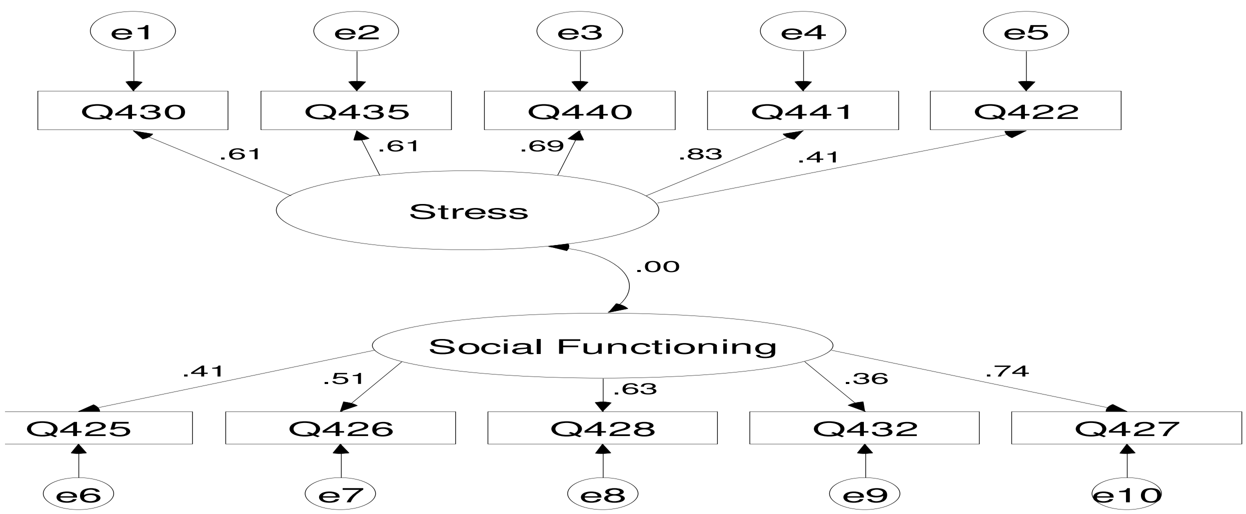
Stress
q422Have you been able to concentrate?
q430 Have you felt constantly under strain?
q435 Have you found everything getting too much for you?
q440 Have you been losing sleep because of worry?
q441: Have you been feeling nervous and strung-up all the time?

Social Functioning:
q425: Have you been managing to keep yourself busy and occupied?
q426: Have you been getting out of the house as much as usual?
q427: Have you felt on the whole that you were doing things well?
q428: Have you felt that you are playing a useful part in things?
q432 Have you been able to enjoy your normal day-to-day activities as much as usual?

(B) Conduct a confirmatory factor analysis using AMOS by drawing a path analytic representation of the factor model that you hypothesized in selecting the GHQ items. Allow the two factors to be correlated.



(C) Conduct a separate CFA forcing the factors to be orthogonal by constraining the correlation between them to be zero.



(D) Check the estimates by conducting an exploratory factor analysis with Stata. Discuss your model and how well it does or does not agree with the exploratory factor analysis. Decide whether or not to correlate your latent variables in your final model and provide a rationale for your decision. If there is substantial disagreement between the results from your exploratory and confirmatory factor analyses, revise your original hypothesis and refit the models.

```
Principal components/correlation                    Number of obs    =       514
Number of comp.  =        10
Trace            =        10
Rotation: (unrotated = principal)            Rho              =    1.0000

--------------------------------------------------------------------------
Component    Eigenvalue   Difference         Proportion   Cumulative
-------------+------------------------------------------------------------
Comp1        3.15615      1.33894            0.3156       0.3156
Comp2        1.81721      .863474            0.1817       0.4973
Comp3        .953735       .16669            0.0954       0.5927
Comp4        .787045     .0527254            0.0787       0.6714
Comp5        .734319     .0494793            0.0734       0.7448
Comp6         .68484      .110776            0.0685       0.8133
Comp7        .574064      .083477            0.0574       0.8707
Comp8        .490587     .0412505            0.0491       0.9198
Comp9        .449337      .096624            0.0449       0.9647
Comp10        .352713            .            0.0353       1.0000
--------------------------------------------------------------------------

factor  q422 q425 q426 q427 q428 q430 q432 q435 q440 q441, ml factors(2)
(obs=514)
Iteration 0:   log likelihood = -77.071215
Iteration 1:   log likelihood = -58.958158
Iteration 2:   log likelihood = -58.660417
Iteration 3:   log likelihood = -58.625027
Iteration 4:   log likelihood = -58.619132
Iteration 5:   log likelihood = -58.618087
Iteration 6:   log likelihood =   -58.6179
Iteration 7:   log likelihood = -58.617866
Iteration 8:   log likelihood =  -58.61786

Factor analysis/correlation                          Number of obs    =      514
    Method: maximum likelihood                       Retained factors =        2
    Rotation: (unrotated)                            Number of params =       19
                                                     Schwarz's BIC    =  235.838
    Log likelihood = -58.61786                       (Akaike's) AIC   =  155.236

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.57538      1.36793            0.6808       0.6808
        Factor2  |      1.20745            .            0.3192       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(45) = 1191.39 Prob>chi2 = 0.0000
    LR test:   2 factors vs. saturated:  chi2(26) =  115.98 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
            q422 |   0.4689    0.0302 |      0.7792
```

```
        q425 |   0.1275    0.4151 |      0.8114
        q426 |   0.2416    0.4524 |      0.7369
        q427 |   0.4758    0.5805 |      0.4366
        q428 |   0.3471    0.5151 |      0.6142
        q430 |   0.5726   -0.2702 |      0.5992
        q432 |   0.4915    0.1106 |      0.7462
        q435 |   0.6081   -0.2033 |      0.5889
        q440 |   0.6605   -0.1404 |      0.5441
        q441 |   0.7474   -0.2844 |      0.3604
    -------------------------------------------------
```

rotate, promax

```
Factor analysis/correlation                        Number of obs    =      514
Method: maximum likelihood                     Retained factors =        2
Rotation: oblique promax (Horst off)           Number of params =       19
Schwarz's BIC    =  235.838
Log likelihood = -58.61786                     (Akaike's) AIC   =  155.236

--------------------------------------------------------------------------
Factor       Variance   Proportion    Rotated factors are correlated
-------------+------------------------------------------------------------
Factor1        2.38603       0.6308
Factor2        1.78523       0.4719
--------------------------------------------------------------------------
LR test: independent vs. saturated:  chi2(45) = 1191.39 Prob>chi2 = 0.0000
LR test:   2 factors vs. saturated:  chi2(26) =  115.98 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
q422	**0.3792**	0.1824	0.7792
q425	-0.1509	**0.4576**	0.8114
q426	-0.0774	**0.5321**	0.7369
q427	0.0414	**0.7365**	0.4366
q428	-0.0270	**0.6292**	0.6142
q430	**0.6549**	-0.0852	0.5992
q432	**0.3482**	**0.2703**	0.7462
q435	**0.6432**	-0.0067	0.5889
q440	**0.6484**	0.0734	0.5441
q441	**0.8122**	-0.0428	0.3604

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9515	0.5925
Factor2	-0.3077	0.8056

```
Factor analysis/correlation                        Number of obs    =      514
Method: maximum likelihood                     Retained factors =        2
```
Rotation: orthogonal varimax (Horst off)   Number of params =   19

```
Schwarz's BIC    =  235.838
Log likelihood = -58.61786                          (Akaike's) AIC   =  155.236

--------------------------------------------------------------------------
Factor       Variance   Difference       Proportion   Cumulative
-------------+------------------------------------------------------------
Factor1        2.23032      0.67781            0.5896       0.5896
Factor2        1.55251            .            0.4104       1.0000
--------------------------------------------------------------------------
LR test: independent vs. saturated:  chi2(45) = 1191.39 Prob>chi2 = 0.0000
LR test:   2 factors vs. saturated:  chi2(26) =  115.98 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

-------------------------------------------------
Variable   Factor1   Factor2    Uniqueness
-------------+--------------------+--------------
q422    0.4079    0.2333       0.7792
q425   -0.0679    0.4289       0.8114
q426    0.0181    0.5126       0.7369
q427    0.1721    0.7306       0.4366
q428    0.0853    0.6153       0.6142
q430    0.6331    0.0091       0.5992
q432    0.3928    0.3154       0.7462
q435    0.6355    0.0847       0.5889
q440    0.6549    0.1643       0.5441
q441    0.7964    0.0731       0.3604
-------------------------------------------------

Factor rotation matrix

--------------------------------
Factor1  Factor2
-------------+------------------
Factor1   0.8982   0.4396
Factor2  -0.4396   0.8982
```

Summary of Findings:

Factor 1 - Stress				
	Path, correlated variables	CFA, promax	Path, uncorrelated variables	CFA, varimax
Q422	.42	.38	.41	.41
Q430	.82	.65	.61	.63
Q435	.69	.64	.61	.64
Q440	.62	.65	.69	.65
Q441	.61	.81	.83	.80
Q432		**.35**		**.039**

Factor 2 – Social Functioning				
	Path, correlated variables	CFA, promax	Path, uncorrelated variables	CFA, varimax
Q425	.38	.46	.41	.43
Q426	.49	.53	.51	.51
Q427	.76	.74	.74	.73

Q428	.62	.63	.63	.62
Q432	.40	.27	.36	.31

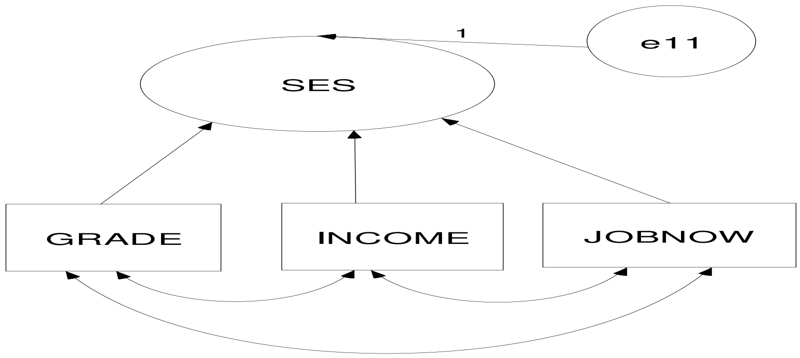
Distress was conceptualized as two variables – Stress and Social Functioning. In general, the path model created of the two factors appears to be similar to the model summary using exploratory and confirmatory factor analytic procedures, though the loadings are slightly higher with the path analytic procedure. With the exception of item q432, which has a fairly equal loading on both Factor 1 and Factor 2, the results of the factory analysis presents a factor structure and loading that are consistent with the regression weights from the path model.

The path models were estimated assuming the two distress variables were correlated, and then uncorrelated (constraining the relationship to 0). Similarly, the factor analysis was conducted using promax and varimax rotations, as the former assumes the factors to be associated, the latter assumed independence. While the individual regression weights are slightly higher when the latent variables are uncorrelated, I have made the decision to correlate them in the final model. This is done for two reasons: (1) the regression weights are not considered to be significantly different from one another; and (2) theoretically, I would expect both latent variables to be associated, given they are both components of a third latent variable – depression.

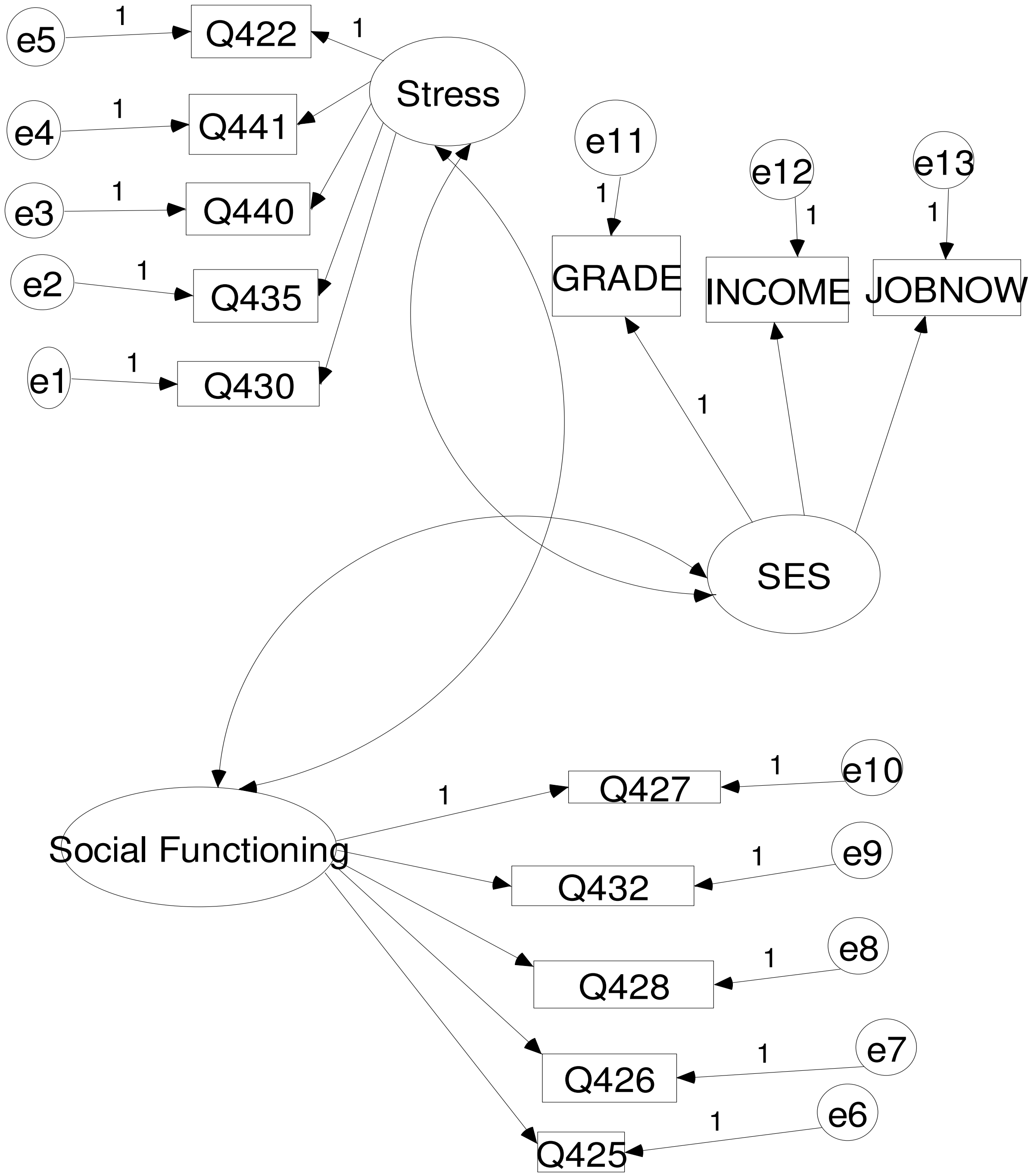
2. Defining socioeconomic status.
(A) Decide how you wish to represent SES in the model, and describe this in words. Be sure to include at least two indicators and decide explicitly whether you are using the scale or index logic for SES.

I would choose to represent SES as an index, comprised of GRADE, INCOME, AND JOBNOW. I believe that GRADE is associated with JOBNOW and INCOME, and that INCOME and JOBNOW are also associated. I would not represent SES as a scale, for both theoretical and empirical reasons. Theoretically, if this were a scale, I would not expect observed variables to change as a function of a change in SES – e.g., past educational attainment would not increase if one's SES increases. Empirically, if it were a scale and SES was a good scale, then all variance in the indicator variables would be accounted for by SES which is inconsistent with the broader conceptual model.

(B) Construct a measurement model for SES, as above.



3. Linking latent variables.
(A) Create a specification that links SES to distress. Include at least two factors of mental distress (these two factors are those from question 1).

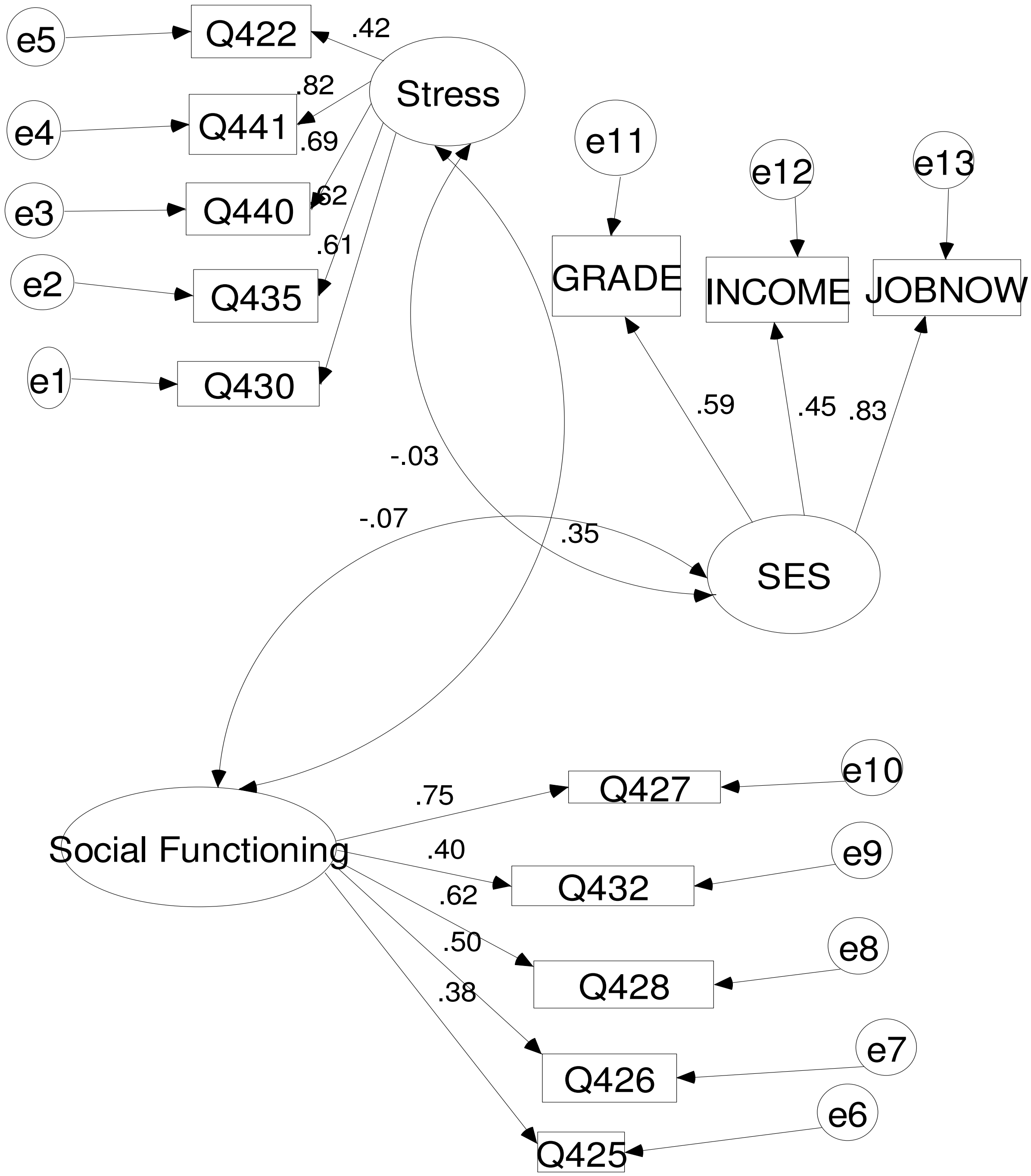


(B) Explain in advance whether your model is identifiable or not.

In first examining the measurement part of the model, the model is identified using the three indictor rule: there is at least one factor, there are at least three indicators per factor, each indicator is only associated with one latent variable, and the errors are not correlated. Whether or not the structural component of the model is identified can

not be determined. The structural component passes the t-rule; there are 6 known parameters, and 6 to be estimated. The model does not, however, pass the null B rule nor the non-recursive rule. As a result, the model might be identified, but additional rules are necessary to determine this for certain.

(C) Estimate the model, then trim and adjust as appropriate.
(D) Print out the diagram with estimates



(E) Discuss your findings. Include in your discussion your evaluation of the fit of the model.

SES was added to the model not as an outcome, but as an associated variable. This was done because, conceptually, neither measure of distress (stress and social functioning) was thought to predict SES but, they are all thought to be associated. In examining the model, one sees that the association between the measures of distress is relatively strong (.35) but the associations between each measure of distress and SES is weak (-.03 and -.07). Conceptually, the negative association between distress and SES is difficult to interpret, as it would suggest that as one's social functioning improve, their SES declines.

Regarding the overall fit of the model, the CMIN for the model is 258.864, p<.000. A significant CMIN is not unexpected, given the large sample size. Examining the BIC, and other fit statistics below:

Model	AIC	BCC	BIC	CAIC
Default model	316.864	318.492	439.889	468.889
Saturated model	182.000	187.106	568.042	659.042
Independence model	1501.400	1502.129	1556.548	1569.548

one finds that the default model run fits the data better than the Saturated model, indicating that the model is more parsimonious – a better fit to the data with fewer parameters estimated.

4. Stratification by gender.
(A) Evaluate whether the measurement model in steps 1 and 2 are equivalent in men and women. Isolate the differences to a limited number of parameters, if possible, and statistically test for the effect of allowing them to vary between groups.

I ran an unconstrained model for men and women, and then ran a second model in which all of the parameters for the measurement model were constrained – i.e., the regression weights for men and women were set to be equal to one another. The CMIN for the default model 308.987, p<.000, while that of the constrained model was 582.152, p<.000. In comparing the two models, they are statistically significantly different from one another, CMIN = 273.165, p<.000. This suggests that the measurement models may be different for men and women. Given the statistically significant difference found between the models, I compared the regression weights for both males and females under the unconstrained models. The results are present below:

	Unconstrained Males	Unconstrained Females
Q430← Stress	.582	.628
Q435← Stress	.568	.635
Q440← Stress	.688	.689
Q441← Stress	.798	.855
Q422← Stress	**.318**	**.474**
Q432←Social Functioning	**.281**	**.430**
Q426←Social Functioning	**.356**	**.633**
Q427←Social Functioning	**.742**	**.725**
Q428←Social Functioning	**.740**	**.560**
Q425←Social Functioning	**.423**	**.380**

Regression weights are bolded if they are thought to be different enough from each other to warrant being unconstrained in the model – i.e., they are different for men and women. In general, the Social Functioning variable seemed to be different for men and women. As a result, all of these parameters comprising this model were left unconstrained, in addition to one item comprising the stress measure. The remaining 4 items were constrained, and set equal to one another for men and women. When comparing the partially constrained model to the default (unconstrained) model, the CMIN for the unconstrained model was not statistically significantly

different from the unconstrained model. This suggests that the partially constrained model would perform just as well as the unconstrained model, with less parameters to be estimated.

(B) Evaluate whether the structural equation model in step 3 is equivalent in men and women in a similar manner.

As in the question above, I first ran an unconstrained model, including the structural component to the model (i.e., the hypothesized associations between SES and the measures of distress). The CMIN for the unconstrained model was 355.799, $\underline{p}<.000$. The table below presents the regression weights for males and females on the unconstrained model:

	Unconstrained Males	Unconstrained Females
Q430← Stress	.599	.623
Q435← Stress	.587	.636
Q440← Stress	.674	.700
Q441← Stress	.773	.847
Q422← Stress	.353	.481
Q432←Social Functioning	**.310**	**.480**
Q426←Social Functioning	**.350**	**.601**
Q427←Social Functioning	.727	.752
Q428←Social Functioning	**.750**	**.538**
Q425←Social Functioning	**.418**	**.342**
JOBNOW←SES	**.708**	**.847**
GRADE←SES	.626	.586
INCOME←SES	**.275**	**.529**
Social Functioning <-> SES	-.117	-.015
Stress <-> SES	-.002	-.001
Stress<-> Social Functioning	**.395**	**.315**

With regard to the structural component, the associations between SES and distress (the structural component) are similar for men and women, but the association between the two measures of distress is stronger for males. In a new model, the distress and SES measures were constrained for men and women (set equal) while the distress association was not. Additional parameters in bold are those considered to be different for men and women, and therefore should not be constrained. The model was re-run, with those variables in plan text constrained. As seen in the table below, the partially constrained model fits the data as well as the unconstrained model, and is more parsimonious as there are fewer parameters to estimate.

Assuming model Default model to be correct:

Model	DF	CMIN	P	NFI Delta-1	IFI Delta-2	RFI rho-1	TLI rho2
Partially Constrained	6	2.372	.883	.002	.002	-.011	-.013