*Bifactor measurement invariance*

Longitudinal measurement invariance of bifactor models was tested at W1-2 (for the YSR subscales) and at W3-4 (for the ASR subscales).The configural invariance combined W1 and W2 bifactor model provided an adequate fit to the data (χ2=941.58(480), p<.001, CFI=.95, TLI=.94, RMSEA=.03, SRMR=.06). It should be noted that only partial invariance could be tested across W1 and W2 due to differences in the ways in which the alcohol use specific factors were constructed in order to accurately capture changes in substance use across this developmental period. A nested model test did not support metric invariance (χ2=183.68(15), *p*<.001). Similarly, the configural invariance combined W3 and W4 bifactor model provided an adequate fit to the data (χ2=1276.87(674), p<.001, CFI=.96, TLI=.95, RMSEA=.03, SRMR=.08), and a nested model test did not support full metric invariance (χ2=64.64(27), *p*<.001). However, when testing metric invariance, the model CFI only decreased by 0.005 at W1-2 and 0.003 at W3-4, which suggests a nonsignificant decrement in model fit (Cheung & Rensvold, 2002). Taken together, it is likely that large sample sizes in the current study allowed nested tests to be well-powered to detect small trivial differences in factor loadings across models.

*Bifactor model fit*

Additional fit indices for the bifactor models indicated that the factor loadings of a unidimensional externalizing factor at each timepoint would have resulted in biased factor loadings. The values of PUC and ECV for the bifactor models supported the specification of bifactor rather than a unidimensional models of externalizing symptoms (PUC range = .82-.81; ECVs range = .48-.54). *H* values were greater than .70 for the externalizing general factor (*H* range = .86-.90) and alcohol use specific factor (*H* range = .86-.94; see Table 4). These high *H* values indicate that the general externalizing and alcohol use specific factors were well-defined by their indicators, and thus, will likely have more stability across studies. *H* values during early adolescence for the aggressive behavior factor suggested that this latent factor was less-adequately represented in early (*H* = .53) to middle adolescence (*H* = .50). The aggression specific factors were better represented by their items in late adolescence (*H* =. 69) to early adulthood (*H* = .70). *H* values for the drug use specific factors indicated that this factor was better represented in middle adolescence (*H* = 0.67) than in late adolescence (*H* = 0.57) and early adulthood (*H* = 0.56). Low *H* values of specific aggression in early and middle adolescence, and low *H* values of specific drug use in late adolescence and early adulthood indicate that these latent variables were not well-defined by their indicators and, thus, are expected to change across studies.

ω and ωS values indicated that the general externalizing factor and specific factors for alcohol and aggression all had good reliability (ω/ωS>.83). ω and ωS values for the drug use specific factors were lower, constituting acceptable reliability (ωS range =.68-.79). ωH was acceptable across all waves for the general externalizing factor (ωH range=.69-.77), and ωHS ranged from .65 to .85 for the alcohol use specific factors, suggesting that these factors were the predominant sources of variance relative to the other specific factors at each wave. ωHS ranged from .40 to .45 for the drug use specific factors, and from .24 to .45 for the aggression specific factors. It is not surprising that ωHS were lower than ωvalues for these specific factors because ωHS values reflect only variance attributable to each specific factor whereas ωis a function of all common variance. Once partitioning out the variance for the general externalizing factor, very little common variance remained for the aggression and drug use specific factors, and thus, subscale reliability decreased. Factor score determinacies, which reflect the correlation between the factor scores and true latent factor scores, were acceptable and ranged from .72 to .97.