

Validation of the Substance Problem Scale (SPS) to the Rasch Measurement Model, GAIN Methods
Report 1.1.

Kendon J. Conrad
University of Illinois at Chicago

Karen M. Conrad
University of Illinois at Chicago

Michael L. Dennis
Chestnut Health Systems

Barth B. Riley
Chestnut Health Systems

Rod Funk
Chestnut Health Systems

Abstract

Purpose. The purpose of this report is to provide a brief psychometric analysis of the *Substance Problem Scale – past year (SPSpy)* using the Rasch measurement model. The *SPSpy* consists of 16 past-year yes/no items related to any alcohol or drug use disorders, including abuse, dependence, and substance induced health and psychiatric problems.

Methods. Data were analyzed on 7,435 persons who presented for substance problem screening. Rasch analysis included an examination of person and item reliabilities; construct validity including item and person fit statistics; fit group analysis; and differential item functioning (DIF) across subgroups. DIF analysis allowed us to determine if the relative item estimates (i.e., item difficulty estimates) remained invariant across subgroups of persons.

Results. The *SPSpy* performs generally well as a unidimensional measure of substance problems with a good Rasch person internal consistency reliability of .80 and an item reliability of 1.00. The Cronbach's alpha is .90. The persons' responses generally conformed to the expectations of the Rasch model. Of the 16 items in *SPSpy*, significant DIF contrasts (i.e., $> .5$ SD=.51 logits) occurred in 5 items for males vs. females, 8 items for youth vs. adults, 5 items for race when using African American as the reference group, and 15 items for primary drugs when using alcohol as the reference group. Using the criterion of .75-1.33 MNSQ for both infit and outfit, the two most misfitting items were "You tried to hide that you were using alcohol or drugs" and "Your alcohol or drug use caused you to have repeated problems with the law." The misfitting item on *hiding use* had a significant DIF in three out of the four contrasts. The misfitting item on *problems with the law* had significant DIF in all four contrasts that were analyzed. In terms of fit group analysis, 83.2% of the sample exhibited person infit and outfit that was low or moderate (L/M) and are thus, regarded as fitting the Rasch model expectations well from a clinical perspective. Slightly over ten percent (10.2%) of the persons had a L/M infit and high (HI) outfit and are referred to as Atypical Type 1, where the overall score may under-estimate severity slightly. Less than one percent (.5%) of the persons had HI infit and L/M outfit (called Atypical Type 2) and tended to be moderate scorers who endorsed a few high severity items. The remaining 6.2% of persons had HI infit and HI outfit (called Atypical Type 3) and tended to be valid high scorers who also have an overall measure but that may underestimate severity more than the other groups, i.e., they need to be highlighted to improve clinical interpretation.

Conclusion. The results suggest that youth and adults should be treated as members of different populations whose data cannot be compared without appropriate adjustments, but that the differences by gender and race were minor and tended to balance out across the full scale. The DIF by substance was

complex and confounded with differences by age. In terms of fit group analysis, the vast majority of the persons fit the Rasch model; thus, their measures are interpretable in the normal way, i.e., low measures indicate less and high indicate more of the construct. Persons in atypical groups 1 and 2 may have *slightly lower* measures than they should given their endorsement of higher severity items. This information should be taken into account clinically. The pattern that is most deceptive for the SPS is Atypical Type 3 since the person measures will tend to be *substantially lower* than they should be based on the person's severity levels on the most severe symptoms. More work on construct validity would be helpful to understand these fit groups better and to ensure proper interpretation.

Citation. Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B.B., & Funk (2009). Validation of the Substance Problem Scale (SPS) to the Rasch Measurement Model, GAIN Methods Report 1.1. Chicago, IL: Chestnut Health Systems. Retrieved from [http://www.chestnut.org/li/gain/psychometric_reports/Conrad et al 2009 SPS Report.pdf](http://www.chestnut.org/li/gain/psychometric_reports/Conrad%20et%20al%202009%20SPS%20Report.pdf)

Purpose of this Report

The purpose of this report is to provide a brief psychometric analysis of the *Substance Problem Scale (SPSpy)* using the Rasch measurement model (Rasch, 1960; Bond & Fox, 2007). The Rasch analysis was conducted using *Winsteps* software (Linacre, 2007). The report presents annotated tables and figures to summarize the main points.

Background

The *SPSpy* is a scale of the *General Individual Severity Scale (GISS)*. The GISS in turn is part of the larger *Global Appraisal of Individual Needs (GAIN)* which is a standardized biopsychosocial instrument that integrates research and clinical assessment for people presenting to substance abuse treatment or other behavioral health treatment (Dennis, Chan, & Funk, 2006).

The *SPSpy* is a count of past-year symptoms related to any alcohol or drug use disorders, including abuse, dependence, substance induced health and psychiatric problems; it is based on the *DSM-IV-TR* (American Psychiatric Association, 2000) and is associated with increased odds of internalizing and externalizing disorders (Dennis, et al., 2006). More specifically, seven items are based on the *DSM-IV-TR* criteria for substance dependence (tolerance, withdrawal, loss of control, inability to quit, time consuming, reduced activity, continued use in spite of medical/mental problems), four items for substance abuse (role failure, hazardous use, continued use in spite of legal problems, continued use in spite of family/social problems), two items for substance-induced disorders (health and psychological), and three items for lower severity symptoms commonly used in screeners (hiding use, people complaining about use, weekly use). The latter five items are not used in diagnosis, but help improve the ability of *SPSpy* to work as a dimensional measure of severity.

SPSpy Items

The *SPSpy* consists of 16 items that assess “recency” of symptoms of substance related problems: The “recency” rating can be used to make symptom counts for the past month, year, or lifetime. In this analysis, we analyzed only the past year data (scored dichotomously as yes=1, no=0). (Please see Conrad, Dennis, Bezruczko, Funk, & Riley [2007] for an analysis of a four-point “recency” response scale). The scale stem reads as follows: “When was the last time that ...” The item stems, GAIN item numbers, Rasch output item location codes, and item labels are shown in the table below.

Table 1. Scale and Item Information

Item Stem	GAIN Item Number	Rasch Output Item Number	Item Label
<i>Substance Problem Scale</i>			
1. you tried to hide that you were using alcohol or drugs?	S9C	1.	HidingUse
2. your parents, family, partner, co-workers, classmates or friends, complained about your alcohol or drug use?	S9D	2.	Complaints
3. you used alcohol or drugs weekly?	S9E	3.	WeeklyUse
4. your alcohol or drug use caused you to feel depressed, nervous, suspicious, uninterested in things, reduced your sexual desire or caused other psychological problems?	S9F	4.	MentHlth
5. your alcohol or drug use caused you to have numbness, tingling, shakes, blackouts, hepatitis, TB, sexually transmitted disease or any other health problems?	S9G	5.	PhysHlth
6. you kept using alcohol or drugs even though you knew it was keeping you from meeting your responsibilities at work, school, or home?	S9H	6.	RoleFailure
7. you used alcohol or drugs where it made the situation unsafe or dangerous for you, such as when you were driving a car, using a machine, or where you might have been forced into sex or hurt?	S9J	7.	HazardousUse
8. your alcohol or drug use caused you to have repeated problems with the law?	S9K	8.	Despite/Legal
9. you kept using alcohol or drugs even though it was causing social problems, leading to fights, or getting you into trouble with other people?	S9M	9.	Trouble/Fights
10. you needed more alcohol or drugs to get the same high or found that the same amount did not get you as high as it used to?	S9N	10.	Tolerance
11. you had withdrawal problems from alcohol or drugs like shaking hands, throwing up, having trouble sitting still or sleeping, or that you used any alcohol or drugs to stop being sick or avoid withdrawal problems?	S9P	11.	Withdrawal
12. you used alcohol or drugs in larger amounts, more often or for a longer time than you meant to?	S9Q	12.	LossControl
13. you were unable to cut down or stop using alcohol or drugs?	S9R	13.	CantStop
14. you spent a lot of time either getting alcohol or drugs, using alcohol or drugs, or feeling the effects of alcohol or drugs (high, sick)?	S9S	14.	TimeConsuming
15. your use of alcohol or drugs caused you to give up, reduce or have problems at important activities at work, school, home or social events?	S9T	15.	GiveUpActs
16. you kept using alcohol or drugs even after you knew it was causing or adding to medical, psychological or emotional problems you were having?	S9U	16.	DespiteHlth

Data Source

Data on the 7435 cases reported in this paper came from 12 projects/programs including 70 sites from around the United States.

All interviews were conducted by interviewers with three to four days of training followed by rigorous field-based certification procedures. Field interviewers had ongoing supervision by local trainers who were trained and certified by Chestnut staff on the use of the GAIN.

Full details about the *SPSpy* may be obtained at the following:

<http://www.chestnut.org/LI/gain/index.html>

Rasch Analysis

The Rasch measurement model (Rasch, 1960) was chosen for this analysis because it is the only item response theory model that has the desirable scaling properties of linear, interval measurement (Embretson & Reise, 2000). Therefore, Rasch measures are the most valid for mathematical operations, such as correlation and regression analysis, as well for assessing change. Rather than tailor models to fit the data, the Rasch one parameter model fulfills the requirements of fundamental measurement, i.e., linear interval scale (Bond & Fox, 2007), and examines the data, i.e., items and persons, for flaws or problems that are indicated by their failure to fit the model.

Quality control with fit statistics. Rasch analysis provides fit statistics to test assumptions of fundamental measurement (Wright & Stone, 1979). “Fitting the model” simply means meeting basic assumptions of measurement, e.g., high scorers should endorse or get right almost all of the easy items. Once identified, persons and items that “misfit” can then be examined qualitatively to determine the causes of the problems. Problems may include items with confusing wording or items that assess a construct that is different from the principal one being measured, i.e., multidimensionality. Understanding poor fit can lead to improving or dropping items.

The fit of the data to the model is evaluated by fit statistics that are calculated for both persons and items. The following link provides a handy guide to interpreting fit statistics: <http://www.rasch.org/rmt/rmt82a.htm>. The Rasch model provides two indicators of misfit: infit and outfit. The infit is sensitive to unexpected behavior affecting responses to items near the person ability level and the outfit is outlier sensitive. Mean square fit statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright & Stone, 1979). Person fit indicates the extent to which the person’s performance is consistent with the way the items are used by the other respondents. Item fit indicates the extent to which the use of a particular item is consistent with the way the sample respondents have responded to the other items. For this type of analysis, values between .75 and 1.33 MNSQ are considered acceptable (Wilson, 2005). In addition to fit statistics, principal component analysis of residuals is used to examine whether a substantial factor exists in the residuals after the primary measurement dimension has been estimated (Linacre, 1998; Smith, 2002).

Construct Validation

In Rasch analysis the item hierarchy that is created by the item difficulty estimates provides an indication of construct validity (Smith, 2001). The items should form a ladder of low severity symptoms on the bottom to high severity symptoms on the top.

In summary, the advantages of Rasch analysis are that:

- Standard errors differ across item and person measures, e.g., improved estimation of error in extreme measures. (In Rasch terminology the word “measure” is used rather than “score.”)
- Enables shorter measures that are more reliable, e.g., eliminate bad items, and via computerized adaptive testing.
- Facilitates analysis of construct validity
- Enables comparable scoring across different measures, i.e., item and test equating.
- Unbiased estimates of item difficulties can be obtained from non-representative samples.
- Interval scale properties are achieved. How? Probabilities, or log odds, are used.
- Analysis of response category usefulness is enhanced.
- Analysis of person and item characteristics is enhanced through fit statistics.
- Enables analysis of item bias, a.k.a., differential item functioning
- Facets beyond persons and items that affect the measures may be estimated

For references to articles that illustrate the applications noted above, we recommend Conrad & Smith (2004). For a complete treatment of Rasch analysis, we recommend Bond & Fox (2007) which includes a glossary of Rasch measurement terminology. Terminology may also be accessed online via *Rasch Measurement Transactions* located at <http://www.rasch.org/rmt/>. The tables below are output from Winsteps (Linacre, 2007) with annotated explanations and interpretations.

Background Characteristics of the Sample

The data for this analysis come from 7435 respondents who completed the *SPSpy*. The respondents were being screened for substance use disorders. In the previous year, 86% had substance use disorders, 51% had internalizing disorders (e.g., somatic, depression, anxiety, trauma, suicide), 59% had externalizing disorders (i.e., attention-deficit or hyperactivity disorders, and conduct disorders), and 59% had problems with crime or violence. Approximately, 42% were entering residential treatment, and 66% were involved in the criminal justice system.

As shown in the following table, the sample was predominantly under 18 years of age (73%) and male (67%). Almost half were Caucasian (45%), a quarter were African American (26%), and the remainder Hispanic or mixed race. Of the top five primary drugs reported, marijuana was reported by 49% of the sample. The primary drug least often reported was opioids at 5%. Other primary drugs reported included amphetamines (11%), cocaine (11%), and alcohol (20.5%). Almost 3% percent of the sample reported other drugs.

Table 2. Demographic Characteristics of the Sample (N=7435^a)

	Percent	Number
Age, Mean (sd) 19.9 (8.9)		
< 18 years	72.5	5388
≥18 years	27.5	2047
Gender		
Male	67.1	4992
Female	32.7	2437
Race		
African American	25.7	1913
Caucasian	45.2	3360
Hispanic	10.8	806
Mixed/other	17.7	1314
Drug, primary, most severe		
Alcohol	20.5	1527
Amphetamines	11.0	820
Marijuana	49.1	3654
Cocaine	10.9	808
Opiates	5.3	393
Other drug	2.9	214

^a Numbers may not add up to 100% due to missing values

Table 3. Person and Item Reliability

SUMMARY OF 6428 MEASURED (NON-EXTREME) PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	8.5	15.9	.23	.69	1.00	.1	1.04	.1
S.D.	4.3	.7	1.67	.17	.23	.8	.76	.9
MAX.	15.0	16.0	3.12	1.27	1.95	3.2	9.90	3.4
MIN.	1.0	5.0	-3.13	.55	.39	-2.6	.20	-2.0
REAL RMSE	.75	ADJ.SD	1.50	SEPARATION	2.00	PERSON RELIABILITY	.80	
MODEL RMSE	.71	ADJ.SD	1.51	SEPARATION	2.12	PERSON RELIABILITY	.82	
S.E. OF PERSON MEAN = .02								

MAXIMUM EXTREME SCORE: 370 PERSONS
 MINIMUM EXTREME SCORE: 567 PERSONS
 LACKING RESPONSES: 14 PERSONS
 DELETED: 56 PERSONS
 VALID RESPONSES: 99.3%

SUMMARY OF 7365 MEASURED (EXTREME AND NON-EXTREME) PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	8.2	15.9	.08	.84				
S.D.	4.9	.8	2.23	.42				
MAX.	16.0	16.0	4.43	2.18				
MIN.	.0	1.0	-4.43	.55				
REAL RMSE	.96	ADJ.SD	2.01	SEPARATION	2.08	PERSON RELIABILITY	.81	
MODEL RMSE	.94	ADJ.SD	2.02	SEPARATION	2.14	PERSON RELIABILITY	.82	
S.E. OF PERSON MEAN = .03								

PERSON RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .90
 (approximate due to missing data)

SUMMARY OF 16 MEASURED (NON-EXTREME) ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	3421.5	6382.7	.00	.03	1.00	-2.4	1.04	-1.7
S.D.	963.2	22.9	1.01	.00	.19	7.3	.36	7.1
MAX.	5281.0	6418.0	2.14	.04	1.47	9.9	2.00	9.9
MIN.	1406.0	6342.0	-2.08	.03	.78	-9.9	.67	-9.9
REAL RMSE	.03	ADJ.SD	1.00	SEPARATION	29.75	ITEM RELIABILITY	1.00	
MODEL RMSE	.03	ADJ.SD	1.00	SEPARATION	30.80	ITEM RELIABILITY	1.00	
S.E. OF ITEM MEAN = .26								

- The 16 item scale has good reliability of .80.
- Cronbach's alpha is higher (.90) because it estimates extreme scores as measured perfectly, i.e., with no error.
- A separation value of 2.12 gives approximately two separation levels, thus splitting the persons into about 3 groups on the Rasch ruler (Figure 1).

- Good item reliability of 1.00.
- Item separation is high at 30.80 meaning items are placed reliably on the ruler.

Figure 1. Wright Map of Persons (#) and Items

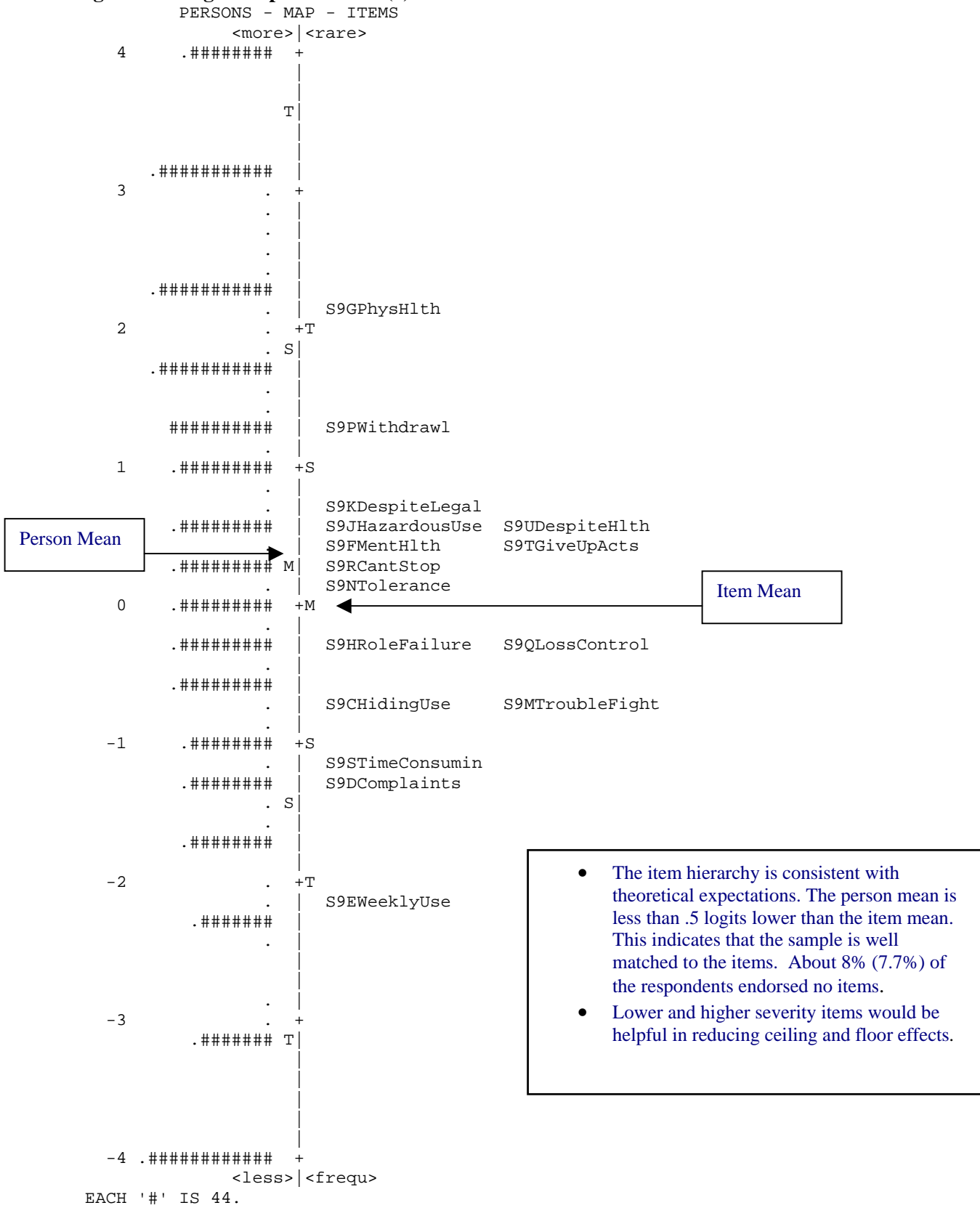


Table 4. Principal Components Analysis of Standardized Residual Correlations for Items

CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --		Modeled
Total raw variance in observations	=	26.8	100.0%	100.0%
Raw variance explained by measures	=	10.8	40.3%	38.2%
Raw variance explained by persons	=	6.1	22.6%	21.4%
Raw Variance explained by items	=	4.8	17.7%	16.8%
Raw unexplained variance (total)	=	16.0	59.7%	100.0% 61.8%
Unexplained variance in 1st contrast	=	1.7	6.4%	10.8%

STANDARDIZED RESIDUAL LOADINGS FOR ITEMS (SORTED BY LOADING)

CON- TRAST	LOADING	INFIT OUTFIT			ENTRY	
		MEASURE	MNSQ	MNSQ	NUMBER	ITEM
1	.66	.65	1.40	1.64	A	8 S9KDespiteLegal
1	.58	-.70	1.09	1.11	B	9 S9MTroubleFights
1	.29	-.73	1.47	2.00	C	1 S9CHidingUse
1	.27	-1.27	1.14	1.38	D	2 S9DComplaints
1	.25	.58	1.06	1.10	E	7 S9JHazardousUse
1	.03	-2.08	.99	1.22	F	3 S9EWeeklyUse
1	-.34	-.33	.86	.75	a	12 S9QLossControl
1	-.33	.54	.84	.75	b	16 S9UDespiteHlth
1	-.32	1.28	.92	.86	c	11 S9PWithdrawl
1	-.31	.41	.78	.67	d	15 S9TGiveUpActs
1	-.31	.49	.90	.86	e	4 S9FMentHlth
1	-.30	.27	.90	.86	f	13 S9RCantStop
1	-.23	.18	.87	.78	g	10 S9NTolerance
1	-.20	-1.11	.86	.73	h	14 S9STimeConsuming
1	-.20	-.31	.84	.74	H	6 S9HRoleFailure
1	-.06	2.14	1.02	1.20	G	5 S9GPhysHlth

- To judge the strength of the measurement dimension, we used the following internal guidelines for variance explained by the measure: $\geq 40\%$ is considered a strong measurement dimension, $\geq 30\%$ is considered a moderate measurement dimension, and $\geq 20\%$ is considered a minimal measurement dimension. The 20% criterion is taken from Reckase (1979).
- 40.3% of the variance is explained by the items, i.e., a strong principal measurement dimension
- 10.8% of the unexplained variance is explained by the 1st residual contrast. Thus, the variance explained by the first factor of residuals supports unidimensionality.
- The table suggests that if there were a second dimension, it would split the measure into “social problems,” such as *S9MTroubleFights* and “health/functioning problems”, such as *S9QLossControl*.
- The PCA was conducted using *Winsteps v. 3.68*

Table 5. Item Misfit Order in Terms of INFIT and OUTFIT Mean Square (MNSQ)

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PT-MEASURE		EXACT MATCH		ITEM	
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%		
1	4562	7351	-.73	.03	1.47	9.9	2.00	9.9	A	.48	.64	67.6	78.2	S9CHidingUse
8	3127	7294	.65	.03	1.40	9.9	1.64	9.9	B	.50	.63	65.8	76.6	S9KDespiteLegal
2	5053	7347	-1.27	.03	1.14	7.5	1.38	7.5	C	.59	.64	78.0	80.7	S9DComplaints
3	5647	7316	-2.08	.04	.99	-1.5	1.22	3.1	D	.62	.63	86.2	85.6	S9EWeeklyUse
5	1775	7347	2.14	.04	1.02	1.1	1.20	3.3	E	.56	.57	82.0	82.5	S9GPhysHlth
9	4497	7285	-.70	.03	1.09	5.5	1.11	3.2	F	.62	.64	76.2	78.1	S9MTroubleFights
7	3202	7280	.58	.03	1.06	3.6	1.10	3.5	G	.61	.63	74.7	76.5	S9JHazardousUse
11	2503	7295	1.28	.03	.92	-4.8	.86	-3.9	H	.63	.61	80.0	78.3	S9PWithdrawl
13	3506	7259	.27	.03	.90	-6.6	.86	-5.4	h	.67	.64	78.8	76.2	S9RCantStop
4	3321	7348	.49	.03	.90	-6.9	.86	-5.0	g	.66	.63	79.0	76.4	S9FMentHlth
10	3604	7285	.18	.03	.87	-8.9	.78	-8.6	f	.68	.64	79.8	76.3	S9NTolerance
14	4871	7290	-1.11	.03	.86	-8.5	.73	-7.1	e	.68	.64	82.9	79.9	S9STimeConsuming
12	4137	7300	-.33	.03	.86	-9.9	.75	-9.0	d	.69	.64	80.7	76.8	S9QLossControl
6	4145	7348	-.31	.03	.84	-9.9	.74	-9.6	c	.69	.64	80.9	76.7	S9HRoleFailure
16	3247	7290	.54	.03	.84	-9.9	.75	-9.6	b	.68	.63	81.3	76.5	S9UDespiteHlth
15	3368	7293	.41	.03	.78	-9.9	.67	-9.9	a	.70	.63	83.0	76.4	S9TGiveUpActs
MEAN	3785.3	7308.0	.00	.03	1.00	-2.4	1.04	-1.7				78.6	78.2	
S.D.	963.9	29.2	1.01	.00	.19	7.3	.36	7.1				5.2	2.6	

Items with high infit mean squares (MNSQ) show a confused or random pattern that is more serious than outfit and reflects that these items are poor indicators of the construct. Two of the infit statistics ("S9ChHidingUse" and "S9KDespiteLegal") were significant using the criterion of .75 to 1.33 (Wilson, 2005). These two items plus "S9DComplaints" were over 1.33 on the outfit statistic. "S9ChHidingUse" was the most misfitting item in terms of outfit (2.00).

Table 6. Most Misfitting Response Strings in terms of Outfit Mean Squares (OUTMNSQ)

MOST MISFITTING RESPONSE STRINGS		PERSON
ITEM	OUTMNSQ	
		6655555444444222111 766555422111776632 55555411
		00932209995547632116611333332860731309234964430155
		55136459883242514839832660557845580156135856169161
		30070265833929771983245287615869420092921725923831
		high-----
1	S9ChHidingUse 2.00	A0.0.....1...11.....
2	S9DComplaints 1.38	C ...0000.0000000.00000.....0.....
3	S9EWeeklyUse 1.22	D 000....0.....0.....0000000000000.....
5	S9GPhysHlth 1.20	E11111111111.111
9	S9MTroubleFig 1.11	F1.....
11	S9PWithdrawl .86	H1.....1...
4	S9FMentHlth .86	g0....1.....
10	S9NTolerance .78	f0.....0.....
14	S9STimeConsum .73	e0.....
		-----low-
		6655555444444222111661176655542211177663295555411
		00932209995547632119832333332860731309234864430155
		55136459883242514833245660557845580156135756169161
		3007026583392977198 287615869420092921 25923831

This table shows the most misfitting items in terms of OUTMNSQ. To the right of the items, we see the persons who unexpectedly endorsed or did not endorse an item given their overall measure. S9ChHidingUse has the highest outfit (2.00) because this very easy item (item measure = -.73 logits) was not endorsed by some persons with high overall measures. For example, Person #5351 (highlighted) did not endorse the item S9ChHidingUse but had a high overall measure of 2.29. The box at the upper left of the table shows the two items with the high outfits.

Table 7. Most Unexpected Responses in Terms of Measure

MOST UNEXPECTED RESPONSES		PERSON
ITEM	MEASURE	
		6655555444444222111 766555422111776632 55555411
		00932209995547632116611333332860731309234964430155
		55136459883242514839832660557845580156135856169161
		30070265833929771983245287615869420092921725923831
		high-----
3	S9EWeeklyUse	-2.08 D 000...0.....0.....0000000000000.....
2	S9DComplaints	-1.27 C ...0000.0000000.00000.....0.....
14	S9STimeConsum	-1.11 e0.....
1	S9ChHidingUse	-.73 A0.0.....1...11.....
9	S9MTroubleFig	-.70 F1.....
10	S9NTolerance	.18 f0.....0.....
4	S9FMentHlth	.49 g0....1.....
11	S9PWithdrawl	1.28 H1.....1...
5	S9GPhysHlth	2.14 E1111111111.111
		-----low-
		6655555444444222111661176655542211177663295555411
		00932209995547632119832333332860731309234864430155
		55136459883242514833245660557845580156135756169161
		3007026583392977198 287615869420092921 25923831

This table shows the items and their measures on the left and the persons with the most unexpected responses on the right.

- For example, Person #4995 has a high measure of 3.12 but did not endorse the easy item S9EWeeklyUse which had a measure of -2.08.
- The large number of 0's for S9EWeeklyUse indicates that the atypical response, i.e., 0, has unexpectedly occurred several times.

Table 8. Person Statistics: Misfit Order

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S. E.	INFIT		OUTFIT		PTMEA CORR.	EXACT OBS%	MATCH EXP%	PERSON
					MNSQ	ZSTD	MNSQ	ZSTD				
125	15	16	3.12	1.07	1.27	.6	9.90	3.1	A-.53	93.8	93.7	125 3 1 0 3 125
134	15	16	3.12	1.07	1.27	.6	9.90	3.1	B-.53	93.8	93.7	134 4 1 0 2 134
1511	1	16	-3.13	1.07	1.27	.6	9.90	3.2	C-.55	93.8	93.7	1511 6 1 1 4 1511
1563	1	16	-3.13	1.07	1.27	.6	9.90	3.2	D-.55	93.8	93.7	1563 6 1 1 2 1563
2317	15	16	3.12	1.07	1.27	.6	9.90	3.1	E-.53	93.8	93.7	2317 3 1 0 2 2317
4118	1	16	-3.13	1.07	1.27	.6	9.90	3.2	F-.55	93.8	93.7	4014 1 1 1 2 4118
4995	15	16	3.12	1.07	1.27	.6	9.90	3.1	G-.53	93.8	93.7	6673 5 1 1 1 4995
5362	1	16	-3.13	1.07	1.27	.6	9.90	3.2	H-.55	93.8	93.7	4140 3 1 0 2 5362
5419	1	16	-3.13	1.07	1.27	.6	9.90	3.2	I-.55	93.8	93.7	4197 3 1 0 2 5419
5465	1	16	-3.13	1.07	1.27	.6	9.90	3.2	J-.55	93.8	93.7	4243 3 2 0 2 5465
5652	1	16	-3.13	1.07	1.27	.6	9.90	3.2	K-.55	93.8	93.7	4430 1 1 0 4 5652
5910	15	16	3.12	1.07	1.27	.6	9.90	3.1	L-.53	93.8	93.7	7218 1 1 1 2 5910
6050	15	16	3.12	1.07	1.27	.6	9.90	3.1	M-.53	93.8	93.7	7358 4 2 1 2 6050
6053	15	16	3.12	1.07	1.27	.6	9.90	3.1	N-.53	93.8	93.7	7361 1 2 1 2 6053
6962	2	16	-2.29	.80	1.51	1.0	7.63	3.4	O-.64	87.5	87.5	5340 3 1 0 4 6962
2059	14	16	2.29	.80	1.51	1.0	7.27	3.3	P-.63	87.5	87.5	2059 1 1 0 4 2059
5356	14	16	2.29	.80	1.50	1.0	6.96	3.2	Q-.60	87.5	87.5	7034 1 2 1 2 5356
3332	2	16	-2.29	.80	1.47	1.0	6.42	3.0	R-.49	87.5	87.5	3228 2 2 0 3 3332
4888	14	16	2.29	.80	1.48	1.0	6.36	3.0	S-.53	87.5	87.5	6566 1 1 1 2 4888
5351	14	16	2.29	.80	1.48	1.0	6.36	3.0	T-.53	87.5	87.5	7029 4 2 1 2 5351
6219	2	16	-2.29	.80	1.33	.7	5.71	2.8	U-.27	87.5	87.5	4597 4 2 0 4 6219
987	2	16	-2.29	.80	1.33	.7	5.70	2.8	V-.27	87.5	87.5	987 3 2 0 4 987
2451	2	16	-2.29	.80	1.33	.7	5.70	2.8	W-.27	87.5	87.5	2451 2 1 0 2 2451
7059	2	16	-2.29	.80	1.33	.7	5.70	2.8	X-.27	87.5	87.5	5437 2 2 0 1 7059
2646	14	16	2.29	.80	1.40	.9	5.60	2.8	Y-.36	87.5	87.5	5899 2 1 1 2 2646
7362	14	16	2.29	.80	1.40	.9	5.60	2.8	Z-.36	87.5	87.5	5740 3 1 0 2 7362
7310	2	16	-2.29	.80	1.21	.6	5.56	2.8	-.16	87.5	87.5	5688 1 1 0 2 7310
1100	14	16	2.29	.80	1.36	.8	5.46	2.7	-.30	87.5	87.5	1100 1 1 0 2 1100
1754	14	16	2.29	.80	1.20	.5	5.24	2.7	-.15	87.5	87.5	1754 1 1 0 3 1754
5093	1	16	-3.13	1.07	1.26	.6	5.20	2.0	-.33	93.8	93.7	6771 5 2 1 1 5093
1382	14	16	2.29	.80	.95	.1	5.11	2.6	.01	87.5	87.5	1382 2 1 0 4 1382
5275	14	16	2.29	.80	.95	.1	5.11	2.6	.01	87.5	87.5	6953 4 2 1 4 5275
6307	14	16	2.29	.80	.95	.1	5.11	2.6	.01	87.5	87.5	4685 2 1 0 4 6307
6368	14	16	2.29	.80	.95	.1	5.11	2.6	.01	87.5	87.5	4746 5 1 0 4 6368
682	15	16	3.12	1.07	1.26	.6	5.10	2.0	-.33	93.8	93.7	682 3 1 0 2 682
693	15	16	3.12	1.07	1.26	.6	5.10	2.0	-.33	93.8	93.7	693 2 1 0 3 693
1138	15	16	3.12	1.07	1.26	.6	5.10	2.0	-.33	93.8	93.7	1138 1 2 0 4 1138

Person misfit is illustrated here where person #1511 has a low measure but unexpectedly endorsed a high severity. Person #4995, has a high measure, but failed to endorse a low severity item. Summary analyses of person fit are provided below.

Table 9. Persons with Most Misfitting Response Strings in Terms of OUTMNSQ

PERSON						OUTMNSQ	ITEM
							1 1 111 1 1
							3241926035467815
						high-----	
125	125	3	1 0 3	125	9.90	A	0.....1
134	134	4	1 0 2	134	9.90	B	0.....1
1511	1511	6	1 1 4	1511	9.90	C1
1563	1563	6	1 1 2	1563	9.90	D1
2317	2317	3	1 0 2	2317	9.90	E	0.....1
4118	4014	1	1 1 2	4118	9.90	F1
4995	6673	5	1 1 1	4995	9.90	G	0.....1
5362	4140	3	1 0 2	5362	9.90	H1
5419	4197	3	1 0 2	5419	9.90	I1
5465	4243	3	2 0 2	5465	9.90	J1
5652	4430	1	1 0 4	5652	9.90	K1
5910	7218	1	1 1 2	5910	9.90	L	0.....1
6050	7358	4	2 1 2	6050	9.90	M	0.....1
6053	7361	1	2 1 2	6053	9.90	N	0.....1
6962	5340	3	1 0 4	6962	7.63	O11
2059	2059	1	1 0 4	2059	7.27	P	00.....1
5356	7034	1	2 1 2	5356	6.96	Q	0.0.....1
3332	3228	2	2 0 3	3332	6.42	R1...1
4888	6566	1	1 1 2	4888	6.36	S	0..0.....1
5351	7029	4	2 1 2	5351	6.36	T	0..0.....1
6219	4597	4	2 0 4	6219	5.71	U	...1.....1
987	987	3	2 0 4	987	5.70	V	...1.....1
2451	2451	2	1 0 2	2451	5.70	W	...1.....1
7059	5437	2	2 0 1	7059	5.70	X	...1.....1
2646	5899	2	1 1 2	2646	5.60	Y	0.....0.....1
7362	5740	3	1 0 2	7362	5.60	Z	0.....0.....1
						-----low-	
							3211916111417815
							4 2 035 6 1

This table shows the unexpected responses of persons with the highest outfit mean squares (MNSQ)s. Note, Persons #1511, #4995 and #5351 again.

We see that the most misfitting response strings for persons were caused by items, such as, #3 **S9EWeeklyUse** and #5 **S9G PhysHealth**.

Table 10. Persons with the most Unexpected Responses in Terms of Measures

PERSON						MEASURE	ITEM
							1 1 111 1 1
							3241926035467815
						high	-----
125	125	3	1	0	3	125	3.12 A 0.....
134	134	4	1	0	2	134	3.12 B 0.....
682	682	3	1	0	2	682	3.12 .0.....
693	693	2	1	0	3	693	3.12 .0.....
1138	1138	1	2	0	4	1138	3.12 .0.....
1189	1189	2	2	0	4	1189	3.12 .0.....
1241	1241	1	1	0	2	1241	3.12 .0.....
2317	2317	3	1	0	2	2317	3.12 E 0.....
2657	5910	4	1	1	1	2657	3.12 .0.....
2729	2625	3	1	0	2	2729	3.12 .0.....
4442	6120	4	1	1	1	4442	3.12 .0.....
4529	6207	1	1	1	1	4529	3.12 .0.....
4533	6211	1	1	1	1	4533	3.12 .0.....
4983	6661	5	2	1	1	4983	3.12 .0.....
4988	6666	5	1	1	1	4988	3.12 .0.....
4995	6673	5	1	1	1	4995	3.12 G 0.....
5056	6734	5	1	1	1	5056	3.12 .0.....
5242	6920	4	2	1	2	5242	3.12 .0.....
5260	6938	4	2	1	1	5260	3.12 .0.....
5337	7015	1	2	1	2	5337	3.12 .0.....
5910	7218	1	1	1	2	5910	3.12 L 0.....
6050	7358	4	2	1	2	6050	3.12 M 0.....
6053	7361	1	2	1	2	6053	3.12 N 0.....
1100	1100	1	1	0	2	1100	2.29 0.....0.....
1382	1382	2	1	0	4	1382	2.29 0.....
1754	1754	1	1	0	3	1754	2.29 0.....
2059	2059	1	1	0	4	2059	2.29 P 00.....
2646	5899	2	1	1	2	2646	2.29 Y 0.....0.....
4888	6566	1	1	1	2	4888	2.29 S 0..0.....
5275	6953	4	2	1	4	5275	2.29 0.....
5351	7029	4	2	1	2	5351	2.29 T 0..0.....
5356	7034	1	2	1	2	5356	2.29 Q 0.0.....
6307	4685	2	1	0	4	6307	2.29 0.....
6368	4746	5	1	0	4	6368	2.29 0.....
7362	5740	3	1	0	2	7362	2.29 Z 0.....0.....
987	987	3	2	0	4	987	-2.29 V ...1.....1
2451	2451	2	1	0	2	2451	-2.29 W ...1.....1
3332	3228	2	2	0	3	3332	-2.29 R1....1
6219	4597	4	2	0	4	6219	-2.29 U ...1.....1
6962	5340	3	1	0	4	6962	-2.29 O11
7059	5437	2	2	0	1	7059	-2.29 X ...1.....1
7310	5688	1	1	0	2	7310	-2.29 1
1511	1511	6	1	1	4	1511	-3.13 C1
1563	1563	6	1	1	2	1563	-3.13 D1
4118	4014	1	1	1	2	4118	-3.13 F1
5093	6771	5	2	1	1	5093	-3.13 1.
5362	4140	3	1	0	2	5362	-3.13 H1
5419	4197	3	1	0	2	5419	-3.13 I1
5465	4243	3	2	0	2	5465	-3.13 J1
5652	4430	1	1	0	4	5652	-3.13 K1
						low	-----
							3211916111417815
							4 2 035 6 1

This table shows the unexpected responses of persons. Note, Persons #4995, #1511, and #5351 again.

Table 11. Summary of Category Structure

SUMMARY OF CATEGORY STRUCTURE. Model="R"

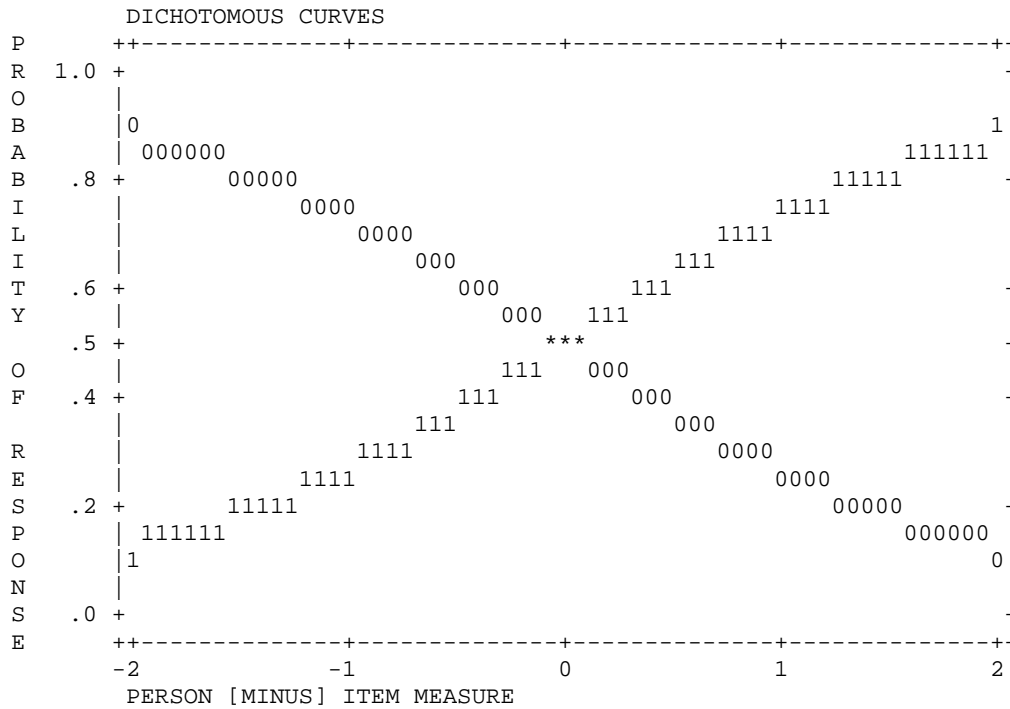
CATEGORY LABEL	OBSERVED SCORE	OBSERVED COUNT	OBSERVED %	OBSVD AVRGE	SAMPLE EXPECT	INFIIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	COHERENCE C->M	ESTIM DISCR
0	0	47379	46	-1.06	-1.06	1.00	1.08	78%	74%	0
1	1	54744	53	1.35	1.35	.99	1.00	78%	81%	1
MISSING		725	1	.39						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

M->C = Does Measure imply Category?

C->M = Does Category imply Measure?

53% of all response options were endorsements, i.e., 1 was chosen as the response option.



Differential Item Functioning (DIF) for Age, Gender, Race, and Primary Drug Severity for the SPSpy

As Bond and Fox (2007) note, the Rasch model requires that relative item estimates (i.e., item difficulty estimates) remain invariant across subgroups of persons (e.g., females and males). DIF analysis allows us to examine whether items have significantly different meanings for different groups. The authors suggest that items that show DIF should be investigated to determine what may be inferred about the underlying construct and what that implies about the samples of persons detected. A significant DIF contrast is based on ≥ 0.5 logit difference for all comparisons which is approximately half a standard deviation (SD) for the scale measure (Norman, Sloan, & Wyrwich, 2003; Conrad et al., 2007). (The SD for the *SPSpy* scale = 1.01; therefore, $SD/2 = .50$).

The figures below present easily interpretable graphs of the relationships of the various groups on the *SPSpy* items. Table 11 contains the data that formed these graphs, and it contains the information to compute differences between groups on each item. For example, to get the DIF contrast between males and females on *S9EWeeklyUse*, subtract $-2.24 - (-1.7) = -.54$.

Gender DIF. For males vs. females, there were five significant contrasts. In Figure 2, we can see that it was significantly more difficult for females to endorse *S9EWeeklyUse*, *S9MTroubleFights* and *S9KDespiteLegal*. Likewise, for males it was significantly more difficult to endorse *S9FMentHealth* and *S9PWithdrawal*. While these were clearly different response patterns for these five items by gender, overall the DIF was not particularly strong since only five items were involved and they tended to balance out the overall differences between males and females. Therefore, there would not tend to be bias in the total *SPSpy* measures i.e., total scores.

Age DIF. The *SPSpy* DIF analysis indicated that eight significant DIF contrasts (i.e., $> .5$ SD) occurred for youth vs. adults. As seen in Figure 3, there are much larger differences between youth and adults than between genders, and they tend to be unbalanced. *S9DComplaints*, *S9CHidingUse*, *S9MTroubleFights*, and *S9KDespiteLegal* are all much more difficult for adults to endorse. Concurrently, there are four items (*S9RCantStop*, *S9FMentHealth*, *S9UDespiteHealth*, *S9PWithdrawal*) that are more difficult for youth to endorse than for adults. For example, as we see in the Table 12 above, the difference between youth and adults on *S9KDespiteLegal* is very large $.23 - 1.83$. *DespiteLegal* is simply much more difficult for adults to endorse and this should be reflected in the measure for each adult. However, if adults are pooled with youth, their measures will be diluted and they will receive measures that are lower (nearer to $.23$) than they should be. Since there were several items that functioned this way, (e.g., *S9ChHidingUse*, *S9DComplaints*, which also had the highest infit mean squares) the tendency in the *SPSpy* is for adults to receive lower scores/measures than they should. This preponderance of highly significant DIF items favoring adults means that adults will tend to get lower measures than they should. In other words, adults will find some items, those involving trouble and legal matters, much harder to endorse than will youth. (See Conrad, Dennis, Bezruczko, Funk, Riley [2007] for a complete treatment of this issue).

Race DIF. As seen in Figure 4, we did not observe much DIF among the five racial/ethnic groups. Using African Americans as the reference group vs. Caucasians, Latinos, and others combined, there were four significant contrasts. Of note was the difference for African Americans whereby *S9CHidingUse* and *S9MTrouble/Fights* were more difficult to endorse. Also, it was more difficult for African Americans to endorse *S9KDespiteLegal* and easier for them to endorse *S9RCantStop*.

Primary Drug Severity DIF. Figure 5 displays the DIF results for primary drug severity. Using alcohol as the reference group, there were 15 out of 16 items that had at least one significant DIF contrast. With six different types of drugs, it appears that opiates and cocaine have the most severe effects. In

other words, persons with opiates and cocaine as the primary drug find the mid and higher severity items to be somewhat easier to endorse. However, the interactions here are quite complex and are the subject of a paper in progress that treats these issues extensively (Dennis, Conrad, & Chan, 2008). This paper indicates that severity hierarchies do differ by drug, but that most of the variation in severity is actually due to the drugs themselves.

Figure 2. SPS Gender DIF

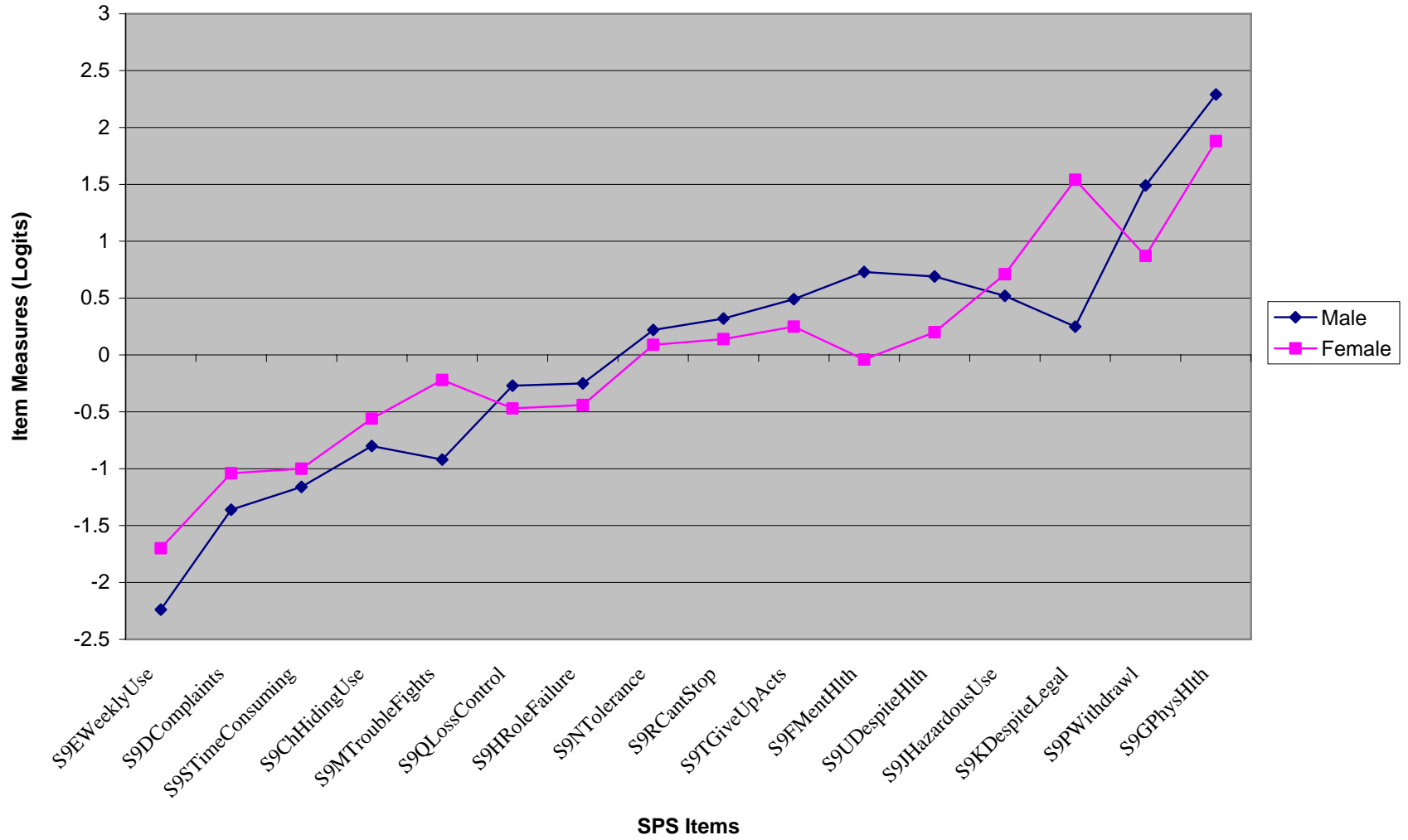


Figure 3. SPS Youth vs. Adult DIF

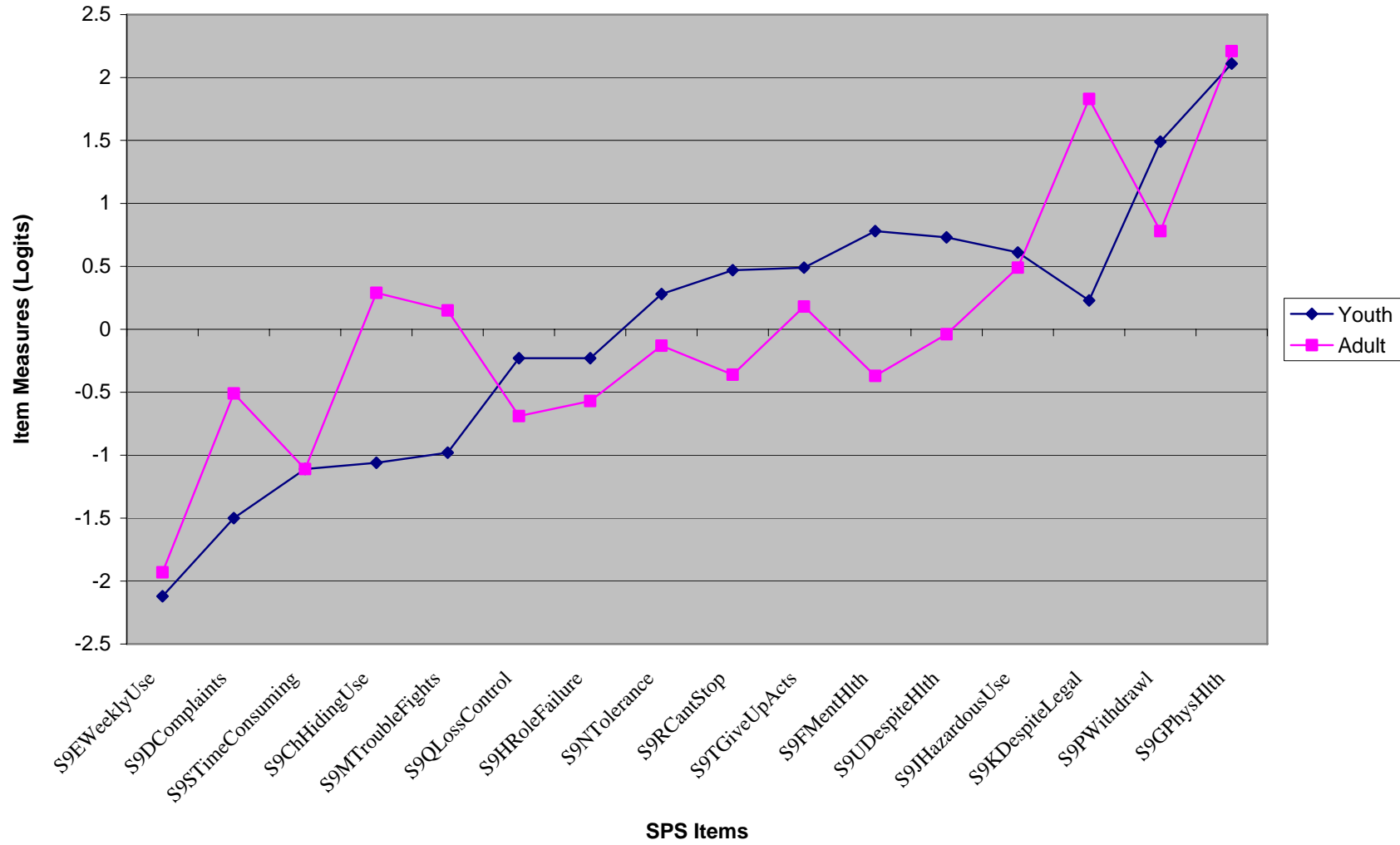


Figure 4. SPS Race DIF

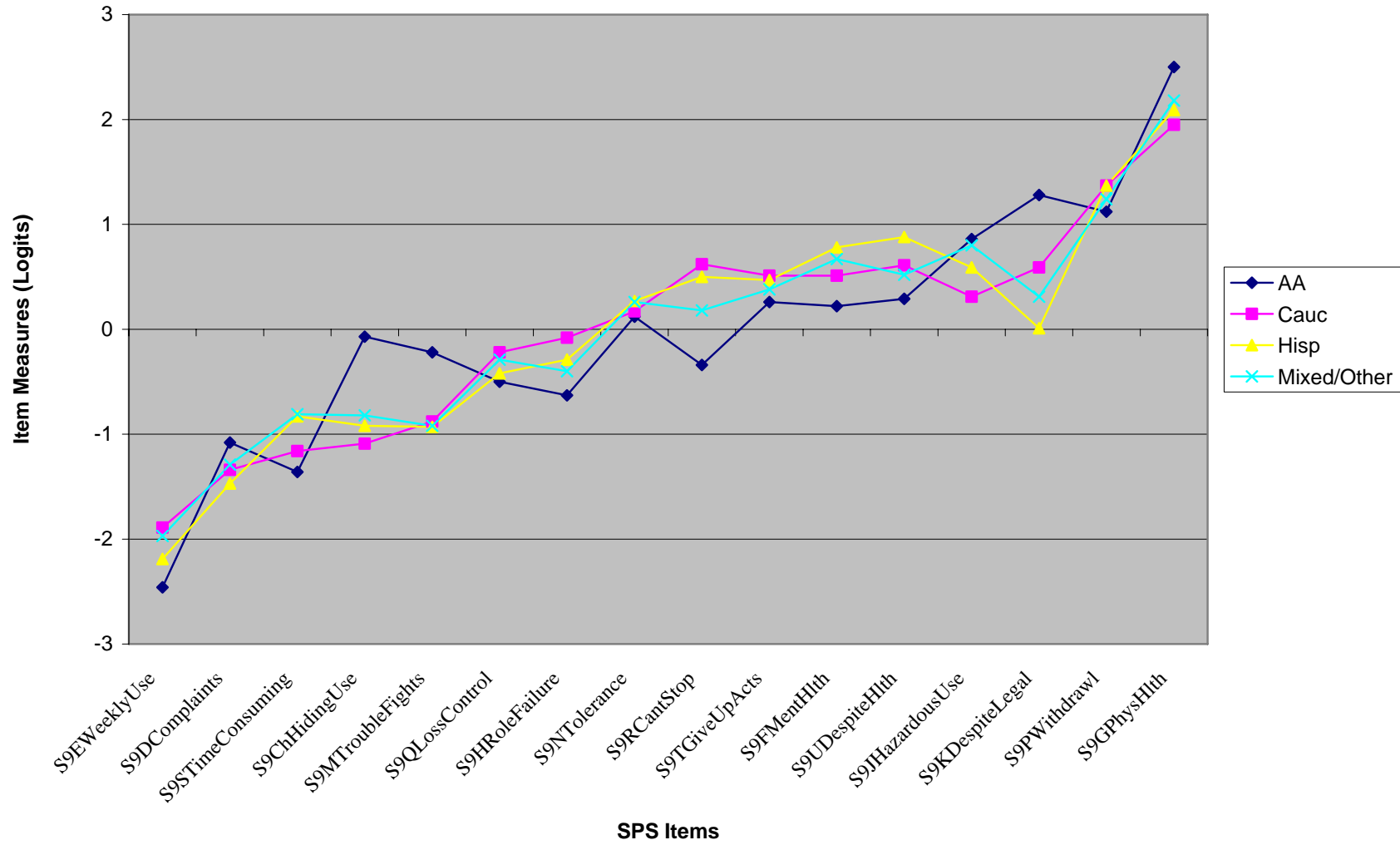


Figure 5. SPS Primary Drug Severity DIF

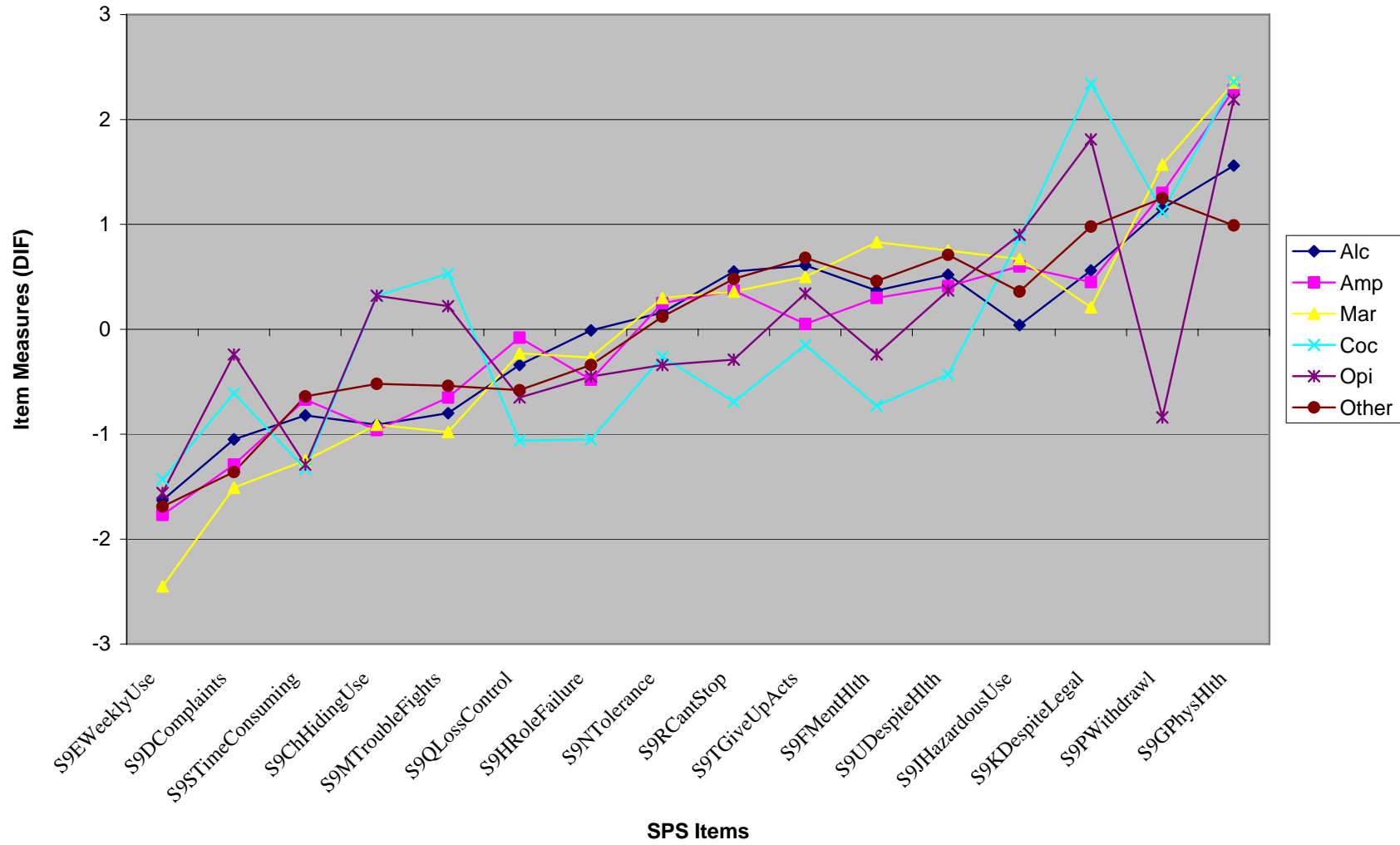


Table 11. *SPSpy* Item Measures by Demographic Groups (Items Listed in Severity Order)

Entry	ITEM	Male	Female	Youth	Adult	AA	Cauc	Hisp	Mixed/ Other	Alc	Amp	Mar	Coc	Opi	Other	Measure
3	S9EWeeklyUse	-2.24	-1.7	-2.12	-1.93	-2.46	-1.89	-2.19	-1.97	-1.63	-1.77	-2.45	-1.43	-1.6	-1.69	-2.08
2	S9DComplaints	-1.36	-1.04	-1.5	-0.51	-1.08	-1.34	-1.47	-1.29	-1.05	-1.29	-1.51	-0.61	-0.2	-1.36	-1.27
14	S9STimeConsuming	-1.16	-1	-1.11	-1.11	-1.36	-1.16	-0.83	-0.81	-0.82	-0.67	-1.25	-1.33	-1.3	-0.64	-1.11
1	S9ChHidingUse	-0.8	-0.56	-1.06	0.29	-0.07	-1.09	-0.92	-0.82	-0.91	-0.96	-0.91	0.32	0.32	-0.52	-0.73
9	S9MTroubleFights	-0.92	-0.22	-0.98	0.15	-0.22	-0.88	-0.93	-0.92	-0.8	-0.65	-0.98	0.53	0.22	-0.54	-0.7
12	S9QLossControl	-0.27	-0.47	-0.23	-0.69	-0.5	-0.22	-0.42	-0.29	-0.34	-0.08	-0.23	-1.06	-0.7	-0.58	-0.33
6	S9HRoleFailure	-0.25	-0.44	-0.23	-0.57	-0.63	-0.08	-0.29	-0.4	-0.01	-0.48	-0.27	-1.05	-0.5	-0.34	-0.31
10	S9NTolerance	0.22	0.09	0.28	-0.13	0.12	0.17	0.27	0.26	0.16	0.25	0.3	-0.26	-0.3	0.12	0.18
13	S9RCantStop	0.32	0.14	0.47	-0.36	-0.34	0.62	0.5	0.18	0.55	0.37	0.36	-0.69	-0.3	0.48	0.27
15	S9TGiveUpActs	0.49	0.25	0.49	0.18	0.26	0.51	0.47	0.38	0.61	0.05	0.5	-0.15	0.34	0.68	0.41
4	S9FMentHlth	0.73	-0.04	0.78	-0.37	0.22	0.51	0.78	0.67	0.37	0.3	0.83	-0.73	-0.2	0.46	0.49
16	S9UDespiteHlth	0.69	0.2	0.73	-0.04	0.29	0.61	0.88	0.52	0.52	0.41	0.75	-0.43	0.37	0.71	0.54
7	S9JHazardousUse	0.52	0.71	0.61	0.49	0.86	0.31	0.59	0.8	0.04	0.6	0.67	0.87	0.9	0.36	0.58
8	S9KDespiteLegal	0.25	1.54	0.23	1.83	1.28	0.59	0.01	0.31	0.56	0.45	0.21	2.34	1.81	0.98	0.65
11	S9PWithdrawl	1.49	0.87	1.49	0.78	1.12	1.37	1.37	1.24	1.15	1.3	1.57	1.12	-0.8	1.25	1.28
5	S9GPhysHlth	2.29	1.88	2.11	2.21	2.5	1.95	2.09	2.18	1.56	2.28	2.35	2.36	2.19	0.99	2.14

Person Fit Group Analysis

The purpose of the person fit group analysis is to illustrate and interpret the expected and unexpected patterns of raw scores in terms of the expectations of the Rasch model. This information should enable us to interpret certain unusual patterns of scores more appropriately, e.g., low scorers who are actually at high risk because of their binge drinking. This type of analysis should inform the interpretation of Rasch measures and enable better treatment decisions. We alert the reader that these charts present raw score p-values (higher proportions endorsing the items are higher) so they are upside down from the typical Rasch charts where more rare is higher.

Figure 6 displays the numbers and percentages of persons in each fit group. In Figures 7-10 below, we present the four possible person fit patterns using Rasch person fit statistics (Wright & Stone, 1979) where ≤ 1.33 mean square on both infit and outfit is low or moderate (L/M) fit (Wilson, 2005). We are regarding this as good fit from a clinical perspective, though we recognize, as we noted earlier, that some would say that very low values, e.g., $< .75$ would be over-fitting. Infit or outfit values above 1.33 are regarded as high or poor fitting patterns (HI).

Therefore, in Figure 7, L/M on infit and L/M on outfit would be a pattern that is consistent with Rasch model expectations, i.e., good fit, and 83% of the persons were in this fit group. The solid lines represent the actual item (dichotomous 0/1 categories) response means for each fit group, and the dashed lines represent the item means over all persons. The red vertical lines indicate the difference between the overall means and fit group means. In Figure 7, the typical group fits the pattern very closely.

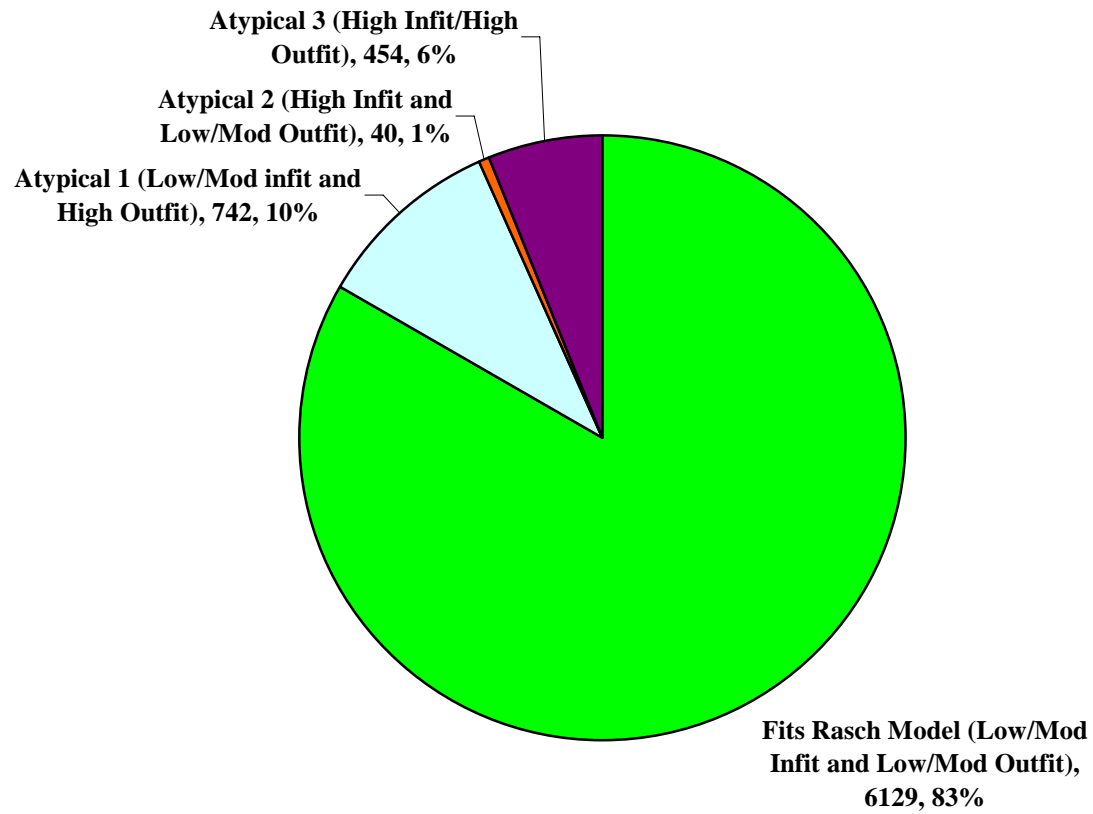
In Figure 8, the L/M infit and HI outfit group (10%) is called Atypical Type 1, where the overall score may underestimate severity since these tend to be people who are high on severe symptoms (higher risk) but are unexpectedly low on “weekly use,” “complaints,” and “hiding use” (which are lower in severity). We interpret this as meaning that this fit group will have slightly lower measures than they should given the high severity items that they tend to endorse.

In Figure 9, the HI infit and L/M outfit group (1%) is called Atypical Type 2 who tend to endorse more low and high severity items than average and less likely than average to endorse three moderate severity items: loss of control, role failure, and giving up other activities. As was the case with Atypical Type 1, these persons will tend to have measures that are slightly lower than they should be given their endorsement of high severity items. Clinically, this group appears to be higher functioning than you would expect given the number and severity of other items they endorse.

In Figure 10, the HI infit and HI outfit group (6%) called Atypical Type 3 who tend to endorse less than average on low severity items (e.g., weekly use, complaints about use, hiding use) and more than average on high severity items (e.g., withdrawal, substance induced health and mental health problems, and hazardous use). These persons will tend to have measures that are significantly lower given their endorsement of high severity items. Clinically, this group appears to be having episodic or binge use that is leading to significant problems, and they may still be early in their stage of addiction (e.g., early and episodic use of crack, opiates and/or methamphetamines).

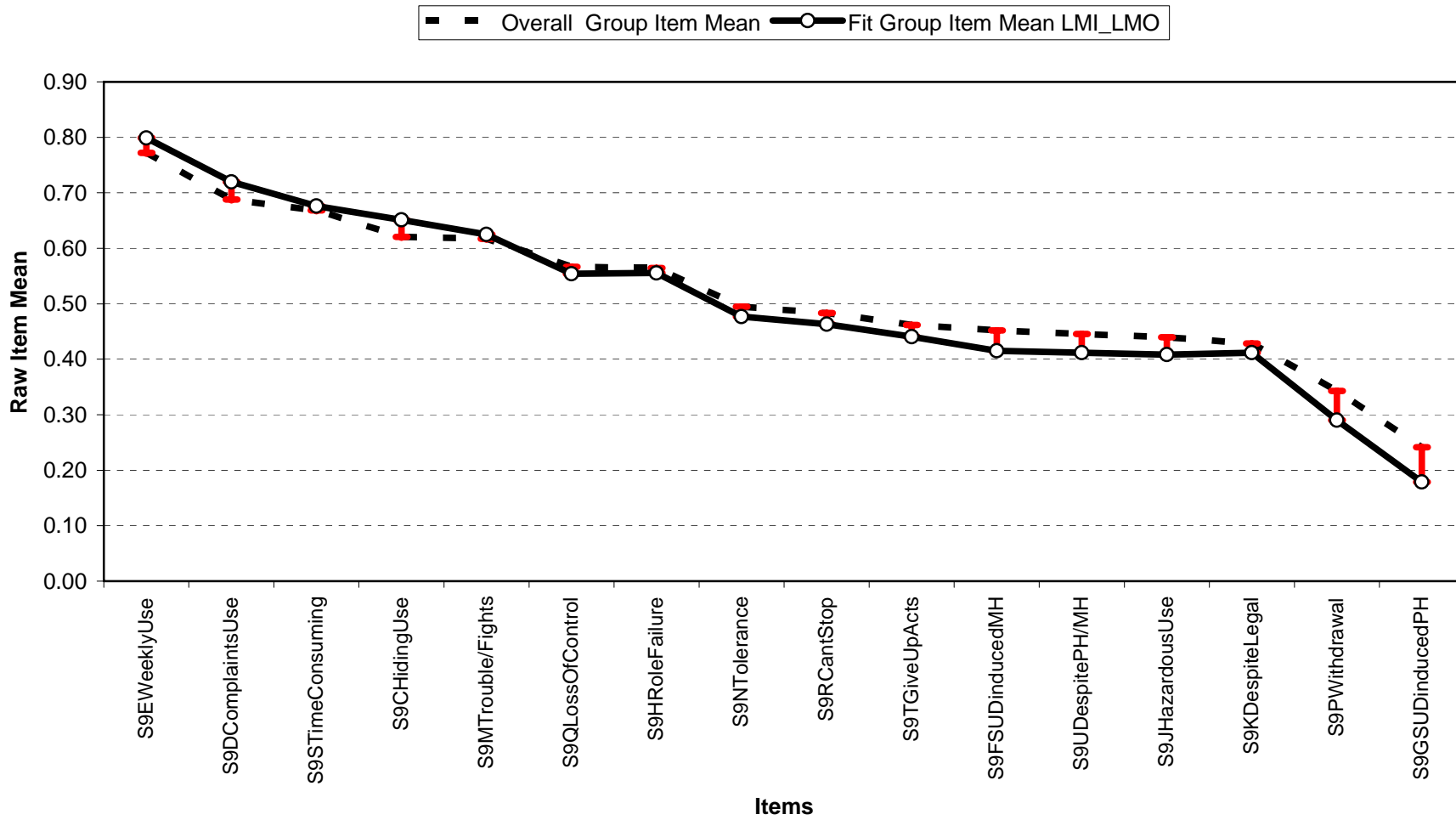
In summary, the pattern that is most deceptive for the SPS is Atypical Type 3 since the person measures will tend to be substantially lower than they should be based on the person’s severity levels on the most severe symptoms. Atypical Types 1 and 2 will also tend to be somewhat deceptive since, even though they will tend to have moderate or high measures, their actual measures should be even higher because they tend to endorse the most severe symptoms.

Figure 6. SPS Fit Group Pie Chart

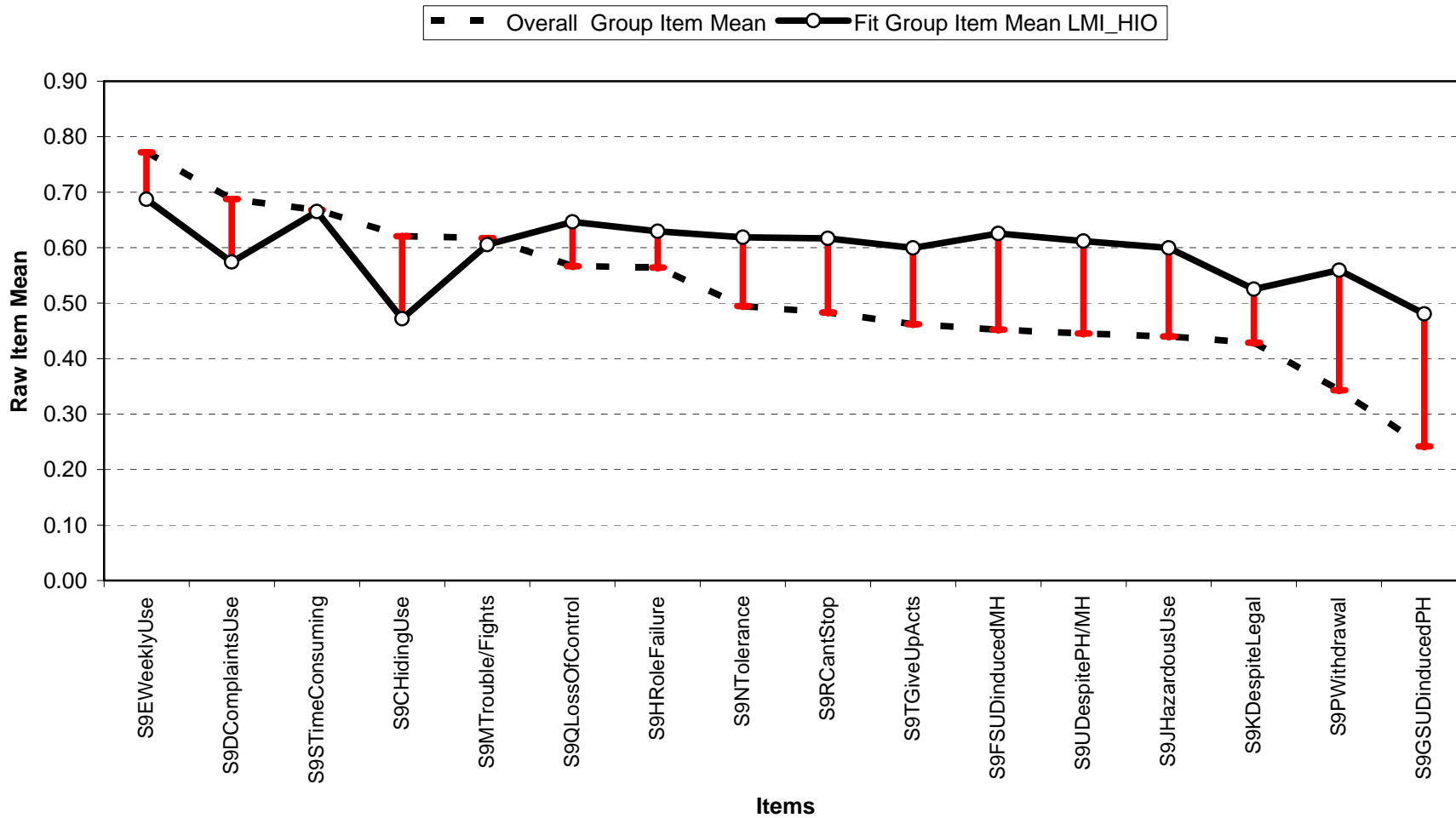


Fits Rasch Model (Low/Mod Infit and Low/Mod Outfit)	Atypical 1 (Low/Mod infit and High Outfit)
Atypical 2 (High Infit and Low/Mod Outfit)	Atypical 3 (High Infit/High Outfit)

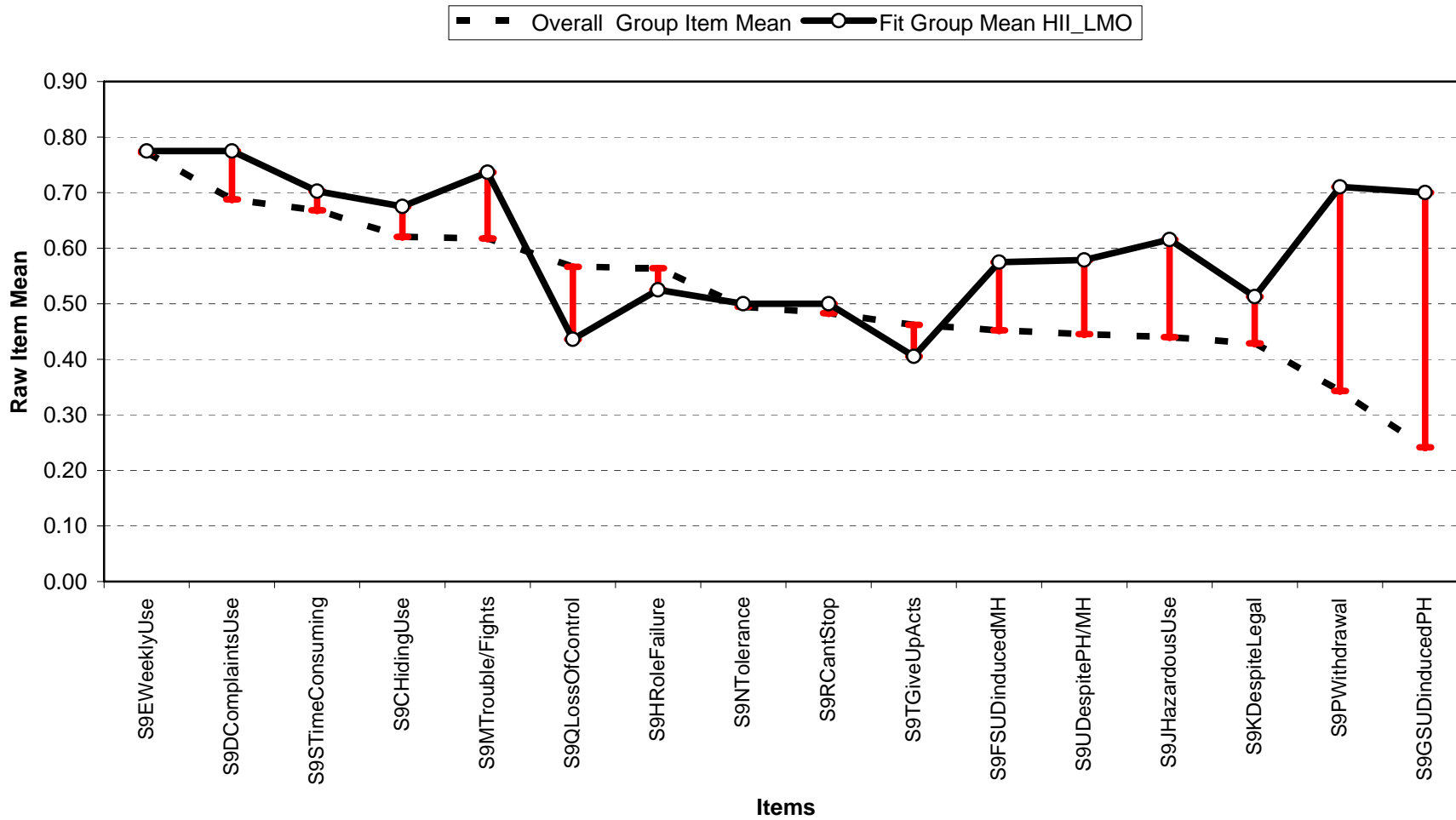
**Figure 7. SPS Overall Group vs. Fit Group Means: Fits Rasch Model (LMLM) (n=6129; 83%)
(Low/Mod on Infit and Low/Mod on Outfit)**



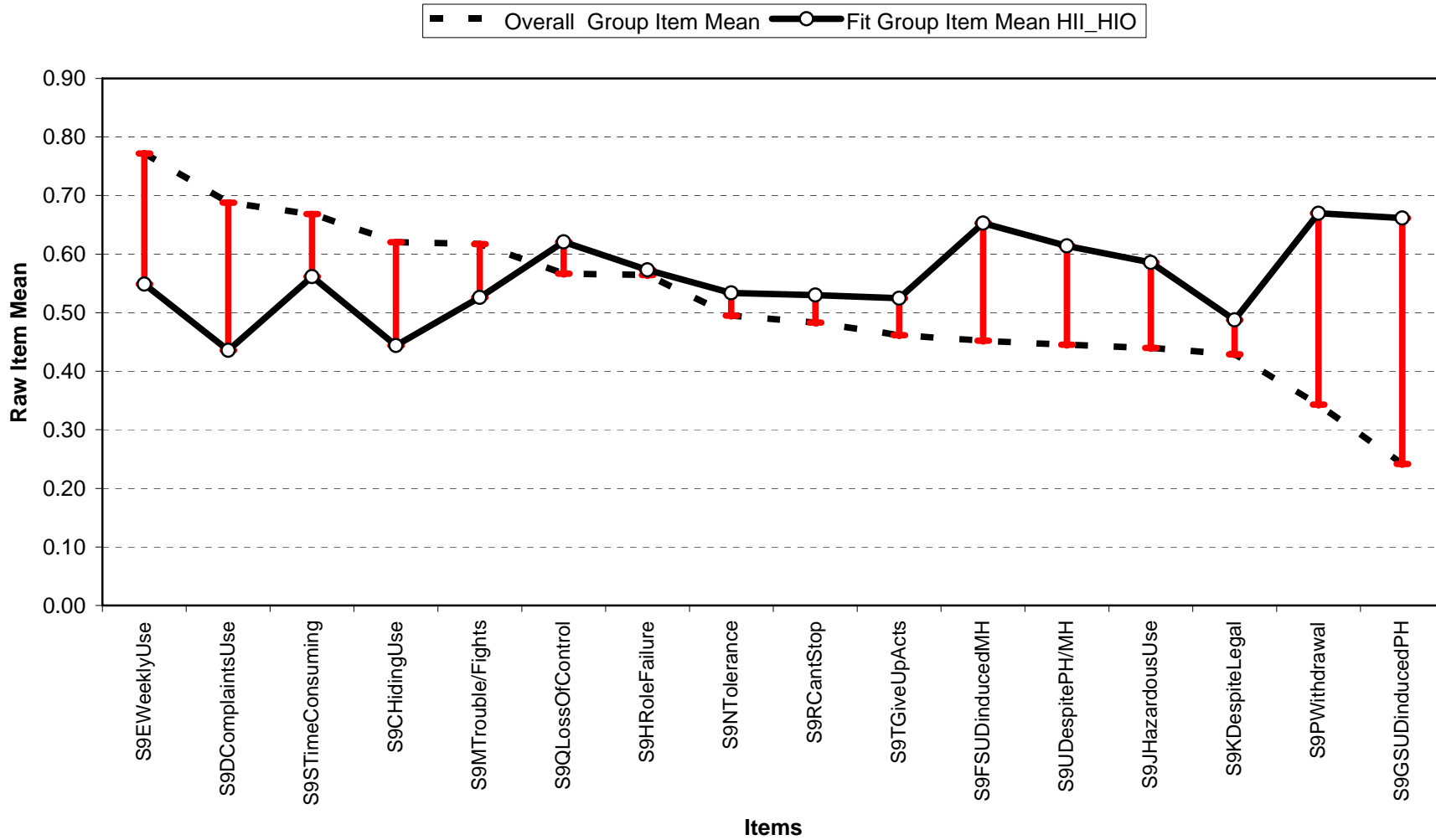
**Figure 8. SPS Overall Group vs. Fit Group Means: Atypical 1 LMHI (n=742; 10%)
(Low/Mod on Infit and High on Outfit)**



**Figure 9. SPS Overall Group vs. Fit Group Item Means: Atypical 2 (HILM) (n=40; 1%)
(High on Infit and Low/Moderate on Outfit)**



**Figure 10. SPS Overall vs. Fit Group Item Means: HIHI (n=454; 6%)
(High on Infit and High on Outfit)**



Recommended Actions and Deliberations.

The SPS_{PY} scale is useful in assessing its target construct. However, there are several recommended actions and deliberations that might improve it.

- Ceiling and floor effects could be reduced or removed by adding some lower and higher severity items. This is especially important since the SUD population is the key target. It would also be useful to add more items to ameliorate response sets to certain items, to develop parallel forms, and for use in computerized adaptive testing.
- The study results suggest that in terms of substance abuse assessment, youth and adults should be treated as members of different populations whose data cannot be pooled nor compared without appropriate adjustments.
- With six different types of drugs, it appears that opiates and cocaine have the most severe effects. In other words, persons with opiates and cocaine as the primary drug find the mid and higher severity items to be somewhat easier to endorse. However, the interactions here are quite complex and are the subject of a paper in progress.
- There was clinically significant DIF ($>$ half a standard deviation) in several analyses. While such DIF usually cancels out in the overall scale, it should be taken into account when interpreting items, for tailoring interventions to these groups (especially age), and in some cases corrected when the goal is to compare these groups.
- Further analyses of DIF and its consequences for measurement of subgroups should be considered.
- Persons having both infit and outfit that are low or moderate (L/M) are regarded as fitting well from a clinical perspective. Here, 83.2% of the persons fit the Rasch model; thus, their measures are interpretable in the usual way, i.e., low measures indicate less and high indicate more of the construct.
- All of the atypical groups, Atypical Types 1, 2, and 3, tend to have measures that underestimate the seriousness of their substance problems. We recommend flagging these three groups for clinicians in the evaluation/validity concerns section of the GAIN Recommendation and Referral Summary (GRRS) as:
 - Atypical 1 response pattern on Substance Problem Scale (Relative to total score, higher than expected on more severe substance problem symptoms, lower than expected on “weekly use,” “complaints,” and “hiding use”)
 - Atypical Type 2 response pattern on Substance Problem Scale (Relative to total score, higher than expected on low and high severity symptoms, lower than expected on three moderate severity symptoms: loss of control, role failure, and giving up other activities)
 - Atypical Type 3 response pattern on Substance Problem Scale (Relative to total score, lower than expected on low severity symptoms, e.g., weekly use, complaints about use, hiding use and higher than expected on higher severity symptom, e.g., withdrawal, substance induced health and mental health problems, and hazardous use). Of the three atypical groups, Atypical 3 will have the most misleading under-estimated score.
- More work on construct validity would be helpful to understand these fit groups better and to ensure proper interpretation of measures.

References

- Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. (2nd Ed.) Mahwah, NJ: Erlbaum Associates.
- Conrad, K.J. & Smith, E.V. (2004). International conference on objective measurement: Applications of Rasch analysis in health care. *Medical Care*, 2004; 42 (suppl I) 1-6.
- Conrad, K.J., Dennis, M.L., Bezruczko, N., Funk, R., & Riley, B. (2007). Substance use disorder symptoms: Evidence of differential item functioning by age. *Journal of Applied Measurement*, 8(4), 373-387.
- Dennis, M. L., Chan, Y-F., & Funk, R. R. (2006). Development and validation of the GAIN short screener (GSS) for internalizing, externalizing, and substance use disorders and crime/violence problems among adolescents and adults. *The American Journal on Addictions*, 15, 80-91.
- Dennis, M. L. Conrad, K. J. & Chan, Y. (2008). Variation in *DSM-IV* symptom severity depending on type of drug. Manuscript in progress.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: Erlbaum.
- Linacre, J.M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, 2(3), 266-283.
- Linacre, J. M. (2007). *Winsteps Rasch measurement computer program* (Version 3.64.2) [Software]. Available from <http://www.winsteps.com/>
- Norman, G.R., Sloan, J.A., & Wyrwich, K.W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, 41(5), 582-592.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut. (Republished Chicago: The University of Chicago Press: 1980).
- Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Smith, E.V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, 2(3), 281-311.
- Smith, E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, 3(2), 205-231.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum Associates.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: University of Chicago, MESA Press.

ACKNOWLEDGEMENT: This development of this paper was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) via Westat under contract 270-2003-00006 to Dr. Dennis at Chestnut Health Systems in Bloomington, Illinois using data provided by the following grants and contracts from CSAT (TI-11320, TI-11317, TI-11321, TI-11323, TI-11324, TI-11422, TI-11424, TI-11423, TI-11894, TI-11874, TI-11888, TI-11892, TI-11871, TI-13309, TI-13356, TI-13305, TI-13340, TI-13344, TI-13322, TI-13323, TI-13345, TI-13308, TI-13354, TI-13313, TI-14254, TI-14376, TI-14311, TI-14196, TI-14214, TI-14261, TI-14090, TI-14189, TI-14252, TI-14283, TI-14355, TI-14272, TI-14103, TI-14267, TI-14315, TI-14188, TI-14271, TI-15686, TI-15671, TI-15486, TI-15545, TI-15672, TI-15475, TI-15678, TI-15447, TI-15461, TI-15433, TI-15481, TI-15514, TI-15478, TI-15413, TI-15483, TI-15670, TI-15674, TI-15479, TI-15682, TI-15467, TI-15511, TI-15562, TI-13601, TI-13190, TI-12541, TI-00567; Contract 207-98-7047, Contract 277-00-6500), the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R01 AA 10368), the National Institute on Drug Abuse (NIDA) (R37 DA11323; R01 DA 018183), the Illinois Criminal Justice Information Authority (95-DB-VX-0017), and the Illinois Office of Alcoholism and Substance Abuse (PI 00567). The opinions are those of the author and do not reflect official positions of the contributing project directors or government.

March 20, 2009