Psychometric Properties of the Wisconsin Schizotypy Scales in an Undergraduate Sample: Classical Test Theory, Item Response Theory, and Differential Item Functioning

By: Beate P. Winterstein, Terry A. Ackerman, Paul J. Silvia and Thomas R. Kwapil

Winterstein, B.P., Ackerman, T., Silvia, P.J., & Kwapil, T.R. (2011). Psychometric properties of the Wisconsin Schizotypy Scales: Classical test theory, item response theory, and differential item functioning. *Journal of Psychopathology and Behavioral Assessment, 33*, 480-490.

Abstract:

The Wisconsin Schizotypy Scales are widely used for assessing schizotypy in nonclinical and clinical samples. However, they were developed using classical test theory (CTT) and have not had their psychometric properties examined with more sophisticated measurement models. The present study employed item response theory (IRT) as well as traditional CTT to examine psychometric properties of four of the schizotypy scales on the item and scale level, using a large sample of undergraduate students ($n = 6{,}137$). In addition, we investigated differential item functioning (DIF) for sex and ethnicity. The analyses revealed many strengths of the four scales, but some items had low discrimination values and many items had high DIF. The results offer useful guidance for applied users and for future development of these scales. The items for the Wisconsin Schizotypy Scales are available from Thomas R. Kwapil.

Keywords: Schizotypy | Wisconsin Schizotypy Scales | Classical test theory | Item response theory | Differential item functioning | psychology

Article:

The present study investigated the psychometric properties of four of the Wisconsin Schizotypy Scales—the Magical Ideation Scale (Eckblad and Chapman *1983*), the Perceptual Aberration Scale (Chapman et al. *1978*), the Revised Social Anhedonia Scale (Chapman et al. *1976*; Eckblad et al. *1982*), and the Physical Anhedonia Scale (Chapman et al. *1976*). After reviewing the multidimensional nature and measurement of schizotypy, we discuss past psychometric investigations of these scales. Despite the popularity of the scales, they have rarely been examined using methods other than classical test theory (CTT). To gain insight into the scales' item-level and scale-level features, we conducted a new large-sample analysis involving CTT, item response theory (IRT), and differential item functioning (DIF).

Conceptualization and Measurement of Schizotypy

The neurodevelopmental vulnerability for schizophrenia is expressed across a dynamic continuum of clinical and subclinical characteristics referred to as schizotypy (e.g., Claridge et al. *1996*; Meehl *1990*). Nonpsychotic people with elevated schizotypy exhibit mild and transient forms of the cognitive, emotional, and behavioral features of schizophrenia, and they are at heightened risk for transitioning into schizophrenia-spectrum disorders. Schizotypy (and by extension schizophrenia) is a multidimensional construct (Claridge et al. *1996*). Positive and negative symptom schizotypy are the most consistently replicated factors (Kwapil et al. *2008*). Positive schizotypy and positive symptom schizophrenia are characterized by odd beliefs and unusual perceptual experiences, which in their extreme form manifest as delusions and hallucinations. Negative schizotypy and schizophrenia are characterized by deficits such as affective flattening, anhedonia, social disinterest, and diminished cognitive functioning.

Psychometric Assessment of Schizotypy

A number of psychometric inventories have been developed for assessing schizotypy in nonclinical samples. These relatively inexpensive and noninvasive measures can be used to screen large samples. The Chapmans and their collaborators developed a series of self-report, true-false questionnaires, such as the Perceptual Aberration (Chapman et al. *1978*), Magical Ideation (Eckblad and Chapman *1983*), Physical Anhedonia (Chapman et al. *1976*) and Revised Social Anhedonia (Eckblad et al. *1982*) Scales, referred to here as the Wisconsin Schizotypy Scales (WSS). The four scales are used as a group to capture different aspects of the complex construct of schizotypy, not as competing or interchangeable measurement instruments.

The WSS were developed and revised using Jackson's (*1970*) recommendations for scale development (prior to the advent or widespread use of IRT and DIF methods). The authors developed formal trait specifications that guided the generation of a large pool of candidate items that were neutral regarding gender, age, and social class (Chapman et al. *1976*). Items were pretested, and decisions for retaining, deleting, or revising them were made based on their high item-scale correlation and their low correlation with social desirability and acquiescence. A further goal was to retain items that had low endorsement rates, to be consistent with the low base rate of schizotypic characteristics and to avoid tapping normal personality traits. The assumption was that each scale would have its maximum discrimination power on the high end of the trait continuum. The revision process was iterative with successive testing. The items from the final scales typically were intermixed when administered (Chapman et al. *1994*).

The psychometric properties of the WSS within the CTT framework have been described in the original source articles and in subsequent reviews (e.g., Kwapil et al. *2002*, *2008*). In brief, research has found good evidence for internal consistency, such as internal-consistency coefficients ranging from .79 to .91 (Winterstein et al. *2010*). In a study of temporal stability, test-retest reliability coefficients (across a period of 2 to 3 months) ranged from .63 to .81, and subsequent generalizability analyses showed that time explained minimal variance (less than 1%) in WSS scores (Winterstein et al. *2010*). Confirmatory factor analyses have provided evidence for the scales' structural validity, particularly the division of the scales into broader positive and

negative symptom dimensions (Kwapil et al. *2008*). Evidence for score validity isn't easily summarized, given the extensive use of the scales over the past several decades, but the WSS have been widely used in cross-sectional and longitudinal studies with psychotic patients and psychosis-prone subjects (e.g., Barrantes-Vidal et al. *2009*, *2010*; Kaczorowski et al. *2009*). In longitudinal research, nonpsychotic people with markedly elevated scores on these scales show psychological and physiological deficits similar to those seen in schizophrenia patients and are at heightened risk for developing schizophrenia-spectrum disorders (e.g., Chapman et al. *1994*; Gooding et al. *2005*, *2007*; Kwapil *1998*).

Prior Psychometric Evaluations of the WSS

Many studies have evaluated the psychometric properties of the WSS using CTT methods (for a review, see Kwapil et al. *2008*). However, there have been few investigations of the WSS using IRT models. Graves and Weinstein (*2004*) conducted Rasch analyses of three of the WSS—the Perceptual Aberration, Magical Ideation, and Revised Social Anhedonia Scales—using a sample of 90 college students. For each scale, they found good evidence for unidimensionality and good fit of the data to the Rasch model. Although an important first step, their research has some limitations that motivate the present analyses. First, the study had a relatively small sample size ($n = 90$). Although there aren't firm sample size guidelines for IRT models, a sample size of 90 seems small for providing trustworthy estimates of IRT and CTT parameters. Second, it omitted the Physical Anhedonia Scale, which is an important part of the set of scales. In particular, including the Physical Anhedonia Scale is essential for capturing the broader dimension of negative-symptom schizotypy. Third, their analysis did not consider models beyond the one-parameter Rasch model. A Rasch model is informative, but it seems likely (and was found in the present research) that a two-parameter model will often be better suited for WSS responses.

Finally, Graves and Weinstein (*2004*) did not examine differential item functioning (DIF) in the WSS, although they indicated that it was an important direction for future work. Exploring DIF in the WSS items is important because past work has found several group differences—such as differences between men and women and differences between racial and ethnic groups (Chmielewski et al. *1995*; Kwapil et al. *2002*)—and it is critical to know if these represent true trait differences. Recent generalizability analyses of the scales (Winterstein et al. *2010*) hint at the possibility of DIF in many of the WSS items. Those analyses found that significant variance in the WSS scores was associated with item-by-person interactions. Such interactions imply the presence of DIF, but generalizability theory is not equipped to examine item-level DIF.

The Present Research

In the present research, we evaluated the WSS using both CTT and IRT methods. Using both measurement approaches offers a well-rounded perspective on the scales' psychometric properties. We particularly focused on IRT and DIF analyses because they have yet to receive much attention in the assessment of schizotypy. To build upon past work, we used a large sample ($n = 6{,}137$) and all four of the WSS. Each scale's item-level and scale-level features were

evaluated, and we examined DIF for both gender (men and women) and for racial groups (African-Americans and Caucasians).

Method

Participants and Materials

A total of 6,137 University of North Carolina at Greensboro undergraduate students enrolled in General Psychology classes participated in this study. The sample included 4,664 women (76%) and 1,473 men (24%), and 4,529 were Caucasian (74%) and 1,608 were African American (26%). Students completed the Magical Ideation, Perceptual Aberration, Revised Social Anhedonia, and Physical Anhedonia Scales as part of mass screening sessions for course credit. The research was approved by the university's Institutional Review Board.

Measures

All WSS items use a binary true/false response format; "false" responses were scored as 0 and "true" responses were scored as 1. After reverse-scoring, all items are scored in the aberrant direction, so endorsing an item means endorsing a schizotypic belief or experience. The Perceptual Aberration Scale (35 items) assesses deviant and distorted perceptual experiences, such as experiencing parts of one's body as detached or decaying. The Magical Ideation Scale (30 items) assesses beliefs in invalid and unlikely causality, such as mind-reading and extraterrestrial influence. Together, the Perceptual Aberration and Magical Ideation Scales capture positive-symptom schizotypy. The Physical Anhedonia Scale (61 items) assesses diminished pleasure from physical and sensory experiences, such as a lack of pleasure from nature, music, and sexuality. The Revised Social Anhedonia Scale (40 items) assesses social disinterest, such as a lack of interest in forming friendships and close relationships. The Physical Anhedonia and Revised Social Anhedonia Scales capture negative-symptom schizotypy, although the Revised Social Anhedonia Scale has been shown to overlap somewhat with positive-symptom schizotypy (Kwapil et al. *2008*; Silvia and Kwapil *2010*).

Statistical Method

The CTT statistics included internal consistency (Kuder-Richardson-20 [KR-20], an estimate of internal consistency for binary data that is equivalent to Cronbach's alpha; DeVellis *2012*) for each scale and the difficulty indices (proportion endorsed, or p-values) and discrimination indices (corrected item-total point-biserial correlations) for each item. To determine the appropriate IRT model, we assessed each scale's unidimensionality, considered the necessity of parameters for discrimination and accounting for false positives, and model fit. To determine whether a 1, 2, or 3 parameter model was appropriate, we first assessed if a difficulty parameter and guessing parameter were necessary and then tested for fit of the 3 models. Decisions regarding the discrimination parameter can be made by looking at the point-biserial correlations. If their range is large, then discrimination should be taken into account. To assess if a parameter for false positives (guessing) was appropriate, we investigated if people low on a trait (based on

the total score for a scale) endorsed items with high difficulty values. The assumption is that if they did not endorse high-difficulty items, then those items do not require a parameter to consider false positives. We included 50 people with the highest total score and 50 people with the lowest total score per scale; concerning items, we decided to include the ones with difficulty values (in CTT) lower than .10 for the Perceptual Aberration Scale, the Revised Social Anhedonia Scale, and the Physical Anhedonia Scale, and 5 items with the lowest difficulty values for the Magical Ideation Scale. In addition, we checked for fit of the 1, 2, and 3-parameter models using Reise's (*1990*) dual approach of assessing fit of people and items.

We used BILOG-MG (Zimowski et al. *2003*) for each scale to create the test information functions and the related curves for standard errors. We also applied BILOG-MG to estimate item parameters (difficulty and discrimination). Simultaneous item bias test (SIBTEST; Stout and Roussos *1999*) was used to investigate DIF in this study. We chose this non-parametric approach because it takes minor factors influencing item responses into account and it assumes only monotonicity. SIBTEST applies a regression correction to take trait differences between people into account. People are separated into bins according to this procedure, with one bin holding people with similar trait levels. Then DIF can be detected by looking at differences in item responses for the reference group and focal group based on the total score. The results indicate which group is favored by an item. Concerning a matching subset of items to test DIF, we chose to pick all items except the one to be tested for DIF.

Because of the large sample size overall, and the considerable sample size of the sex and ethnicity reference groups, we choose to interpret the statistic beta-uni instead of the chi-square or SIB uni z-statistics. Beta-uni can be interpreted as an effect size, and Roussos and Stout (*1996*) provided some guidelines under the assumption that the items meet statistical rejection: A-level items display negligible DIF and have a beta-uni value < 0.059; B-level items display moderate DIF and have a beta-uni value between 0.059 and 0.088; and C-level items display large DIF and have a beta-uni value of 0.088 or greater.

Results

Descriptive Statistics for Scales and Scale Correlations

Table 1 displays the descriptive statistics. The total score distributions are positively skewed with low means and narrow standard deviations. The skew reflects the expected distribution of the traits, given the assumption that only a small percentage of people are schizotypic. Furthermore, the intercorrelations of the scales are consistent with past work (e.g., Kwapil et al. *2008*).

Table 1 Descriptive scale statistics, reliabilities, and correlations

	Number of items	Mean	SD	Variance	Skew	Reliability	Magical Ideation	Perceptual Aberration	Social Anhedonia	Physical Anhedonia
Magical Ideation	30	9.36	5.60	31.37	.63	.84	1	.69	.22	−.10
Perceptual Aberration	35	5.81	5.54	30.69	1.83	.88		1	.29	−.03

	Number of items	Mean	SD	Variance	Skew	Reliability	Magical Ideation	Perceptual Aberration	Social Anhedonia	Physical Anhedonia
Social Anhedonia	40	8.53	5.77	33.24	1.23	.84			1	.42
Physical Anhedonia	61	12.99	7.05	49.74	.83	.84				1

n = 6,137. Reliability refers to KR-20 internal consistency coefficients

Classical Test Theory

The internal consistency coefficients for the schizotypy scales were all in the mid .80s (see Table 1). Tables 2, 3, 4, and 5 present the psychometric properties of the single items according to CTT, including discrimination (corrected item-total point-biserial correlations) and difficulty values (percent who endorsed the item). The ranges of the item discrimination values were .09 to .65 (*M* = .48, *SD* = .13) for the Magical Ideation Scale, .21 to 1.00 (*M* = .69, *SD* = .21) for the Perceptual Aberration Scale, .15 to .80 (*M* = .51, *SD* = .17) for the Revised Social Anhedonia Scale, and .00 to .76 (*M* = .40, *SD* = .14) for the Physical Anhedonia Scale. The Perceptual Aberration Scale seems to differentiate between people high and low on the trait best, which suggests that the scale as a whole is a good indicator of the trait. The Physical Anhedonia Scale was the poorest at differentiating between people. The column labeled “Point-Biserial Corr.” in Tables 2, 3, 4, and 5 indicates which items do the best job of discriminating between people high and low on a trait.

Table 2 The Magical Ideation Scale: psychometric properties according to CTT, IRT, and DIF

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
1	.29	0.50	1.03	0.63	.014	.008	
2	.23	0.46	1.47	0.58	.038	.017	
3	.31	0.47	0.97	0.60	.039	−.037	
4	.23	0.09	6.10	0.12	.003	.006	
5	.24	0.51	1.29	0.66	−.088	−.023	W
6	.16	0.59	1.62	0.81	.000	.008	
7	.33	0.29	1.38	0.31	−.057	−.031	
8	.52	0.54	−0.06	0.74	−.044	.014	
9	.28	0.53	1.05	0.68	−.006	−.013	
10	.55	0.56	−0.17	0.82	.060	−.033	M
11	.39	0.44	0.58	0.53	.044	.005	
12	.24	0.65	1.01	0.95	.053	.037	
13	.29	0.60	0.89	0.84	−.014	−.004	

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
14	.34	0.52	0.76	0.69	.016	.072	AA
15	.36	0.41	0.86	0.45	.054	−.071	C
16	.42	0.59	0.32	0.82	−.014	−.094	C
17	.44	0.42	0.31	0.51	.032	.018	
18	.44	0.36	0.43	0.39	.063	.095	M, AA
19	.44	0.41	0.33	0.49	−.169	−.040	W
20	.55	0.39	−0.27	0.45	.060	−.074	M, C
21	.24	0.54	1.23	0.73	−.084	.016	W
22	.23	0.50	1.37	0.64	−.031	.072	AA
23	.27	0.44	1.31	0.53	.030	.070	AA
24	.24	0.44	1.58	0.50	.020	.019	
25	.10	0.63	1.98	0.92	−.077	.035	W
26	.20	0.64	1.27	0.91	−.037	−.047	
27	.23	0.46	1.52	0.57	.118	−.039	M
28	.31	0.62	0.77	0.87	.055	−.067	C
29	.07	0.61	2.23	0.90	−.051	−.004	
30	.47	0.18	0.38	0.19	−.022	−.011	

M men, *W* women, *AA* African American, *C* Caucasian

Table 3 The Perceptual Aberration Scale: psychometric properties according to CTT, IRT, and DIF

	CTT		IRT		DIF		
	P	Point- biserial corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
1	.17	0.67	1.42	0.91	.048	.023	
2	.19	0.73	1.20	1.08	.001	.046	
3	.28	0.52	1.01	0.71	.059	−.005	M
4	.25	0.41	1.48	0.50	.014	−.021	
5	.41	0.21	1.08	0.21	−.009	−.119	C
6	.35	0.43	0.77	0.53	.058	−.010	
7	.23	0.64	1.13	0.86	−.002	.064	AA
8	.08	0.79	1.94	1.07	−.013	−.023	
9	.22	0.35	2.31	0.35	−.017	−.030	

	CTT		IRT		DIF		
	P	Point- biserial corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
10	.08	0.88	1.76	1.31	.027	−.009	
11	.12	0.75	1.64	1.07	−.050	.039	
12	.14	0.53	2.11	0.62	.010	.001	
13	.05	0.91	2.25	1.19	−.017	.002	
14	.06	0.77	2.32	0.96	−.013	−.008	
15	.22	0.50	1.46	0.62	.042	.092	AA
16	.08	0.81	2.08	1.01	−.018	−.008	
17	.40	0.53	0.43	0.79	.022	.087	AA
18	.42	0.57	0.30	0.90	.069	−.077	M, C
19	.09	0.77	1.90	1.01	−.035	−.012	
20	.30	0.29	1.86	0.27	−.027	.016	
21	.08	0.87	1.78	1.28	−.022	.003	
22	.20	0.50	1.68	0.57	−.010	−.039	
23	.08	0.95	1.76	1.46	.011	−.017	
24	.07	0.79	2.20	0.97	−.009	.017	
25	.09	0.82	1.83	1.12	.012	−.025	
26	.04	1.03	2.08	1.50	−.004	.001	
27	.06	0.94	1.92	1.35	−.007	−.021	
28	.13	0.73	1.60	1.00	−.030	.052	
29	.16	0.75	1.37	1.07	−.013	.057	
30	.06	1.02	1.85	1.66	−.004	.002	
31	.07	0.97	1.80	1.54	.015	−.023	
32	.19	0.64	1.41	0.83	−.032	.022	
33	.23	0.64	1.14	0.87	.042	−.094	C
34	.04	1.03	2.14	1.52	−.001	−.015	
35	.14	0.66	1.66	0.87	−.022	−.073	C

M men, *W* women, *AA* African American, *C* Caucasian

Table 4 The revised social Anhedonia scale: psychometric properties according to CTT, IRT, and DIF

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
1	.11	0.70	1.76	0.99	.037	−.049	
2	.07	0.73	2.03	1.04	.010	−.056	
3	.11	0.50	2.27	0.65	.004	.029	
4	.11	0.67	1.82	0.94	.031	−.020	
5	.39	0.43	0.55	0.54	−.055	−.027	
6	.26	0.42	1.42	0.51	.019	.010	
7	.04	0.79	2.49	1.11	−.002	.004	
8	.31	0.34	1.32	0.39	.038	.080	AA
9	.67	0.32	−1.10	0.42	.136	−.028	M
10	.08	0.75	2.02	1.03	.003	.018	
11	.19	0.15	4.87	0.17	−.083	−.064	C
12	.24	0.48	1.36	0.58	−.132	.009	W
13	.17	0.65	1.43	0.91	−.008	.064	AA
14	.17	0.67	1.42	0.95	.045	.037	
15	.14	0.49	2.10	0.59	−.016	−.021	
16	.27	0.36	1.59	0.42	−.179	.106	W, AA
17	.11	0.66	1.90	0.90	−.012	.002	
18	.15	0.47	2.08	0.56	−.117	.055	W
19	.06	0.80	2.12	1.15	−.023	.005	
20	.30	0.48	1.00	0.60	−.025	−.009	
21	.17	0.58	1.62	0.74	−.013	−.062	C
22	.31	0.21	2.02	0.24	.118	−.157	M, C
23	.24	0.32	1.95	0.38	.004	−.095	C
24	.23	0.32	2.10	0.37	.060	−.095	M, C
25	.47	0.41	0.19	0.50	−.008	−.082	C
26	.21	0.71	1.09	1.07	.071	.020	M
27	.05	0.55	2.83	0.72	−.083	.016	W
28	.40	0.44	0.49	0.54	.028	.009	
29	.47	0.42	0.17	0.54	.039	−.097	C
30	.12	0.75	1.58	1.13	.011	.039	
31	.17	0.53	1.71	0.65	−.015	−.024	
32	.26	0.40	1.50	0.48	−.050	.034	
33	.28	0.41	1.32	0.49	.078	.123	M, AA

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
34	.49	0.42	0.06	0.54	.089	−.075	M, C
35	.09	0.58	2.30	0.74	−.019	.020	
36	.11	0.37	3.11	0.43	−.008	−.032	
37	.06	0.74	2.21	1.03	.024	−.016	
38	.17	0.35	2.57	0.40	−.032	−.004	
39	.19	0.42	1.93	0.50	−.077	.093	W, AA
40	.06	0.55	2.71	0.70	−.061	−.009	W

M men, *W* women, *AA* African American, *C* Caucasian

Table 5 The physical Anhedonia scale: psychometric properties according to CTT, IRT, and DIF

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
1	0.21	0.40	1.76	0.52	.028	.092	AA
2	0.22	0.13	6.21	0.12	.129	.051	M
3	0.09	0.58	2.04	0.87	−.019	.037	
4	0.35	0.24	1.36	0.28	−.143	.083	W, AA
5	0.38	0.41	0.66	0.52	−.118	−.496	W, C
6	0.13	0.36	2.79	0.45	−.098	.046	W
7	0.22	0.30	2.27	0.35	.147	.037	M
8	0.62	0.23	−1.15	0.27	−.019	.142	AA
9	0.13	0.43	2.26	0.58	−.035	.079	AA
10	0.19	0.55	1.44	0.77	−.078	.047	W
11	0.79	−0.01	−14.82	0.05	.082	−.009	
12	0.51	0.51	−0.02	0.71	−.234	−.078	W, C
13	0.10	0.43	2.68	0.56	−.136	.102	W, AA
14	0.44	0.32	0.40	0.38	.160	.039	M
15	0.13	0.46	2.15	0.61	−.017	−.023	
16	0.14	0.22	4.51	0.25	.055	.053	
17	0.16	0.20	4.07	0.24	.042	−.066	C
18	0.22	0.42	1.69	0.52	−.119	.074	W, AA
19	0.28	0.47	1.12	0.62	−.042	.080	AA
20	0.08	0.40	3.03	0.54	.027	.020	

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
21	0.14	0.25	4.61	0.24	.055	.013	
22	0.10	0.41	2.82	0.52	−.031	.051	
23	0.15	0.39	2.36	0.48	−.025	−.038	
24	0.12	0.61	1.79	0.93	.021	−.014	
25	0.23	0.23	2.76	0.27	.031	−.122	C
26	0.31	0.31	1.40	0.36	.125	−.056	M
27	0.21	0.37	1.94	0.45	−.014	−.042	
28	0.32	0.34	1.28	0.38	.052	.119	AA
29	0.34	0.52	0.73	0.71	.065	−.100	M, C
30	0.20	0.42	1.75	0.53	−.099	.114	W, AA
31	0.44	0.38	0.36	0.47	−.088	−.355	W, C
32	0.31	0.49	0.96	0.63	.127	−.089	M, C
33	0.08	0.38	3.15	0.50	−.003	−.003	
34	0.08	0.31	3.76	0.41	.017	.054	
35	0.05	0.50	2.96	0.71	−.037	.025	
36	0.09	0.69	1.78	1.15	−.064	.016	W
37	0.08	0.41	3.03	0.52	−.031	.037	
38	0.06	0.32	4.13	0.43	−.001	.013	
39	0.12	0.50	2.04	0.71	.019	−.155	C
40	0.65	0.33	−1.01	0.40	.169	.061	M, AA
41	0.12	0.37	2.90	0.45	.034	−.004	
42	0.06	0.72	2.07	1.26	.003	−.022	
43	0.23	0.38	1.87	0.43	.062	.083	M, AA
44	0.19	0.36	2.35	0.41	.061	.041	M
45	0.23	0.42	1.54	0.53	−.108	.091	W, AA
46	0.07	0.76	1.80	1.39	−.001	.019	
47	0.07	0.62	2.23	0.96	.000	−.018	
48	0.35	0.48	0.72	0.63	−.030	−.200	C
49	0.09	0.35	3.27	0.47	−.007	.009	
50	0.26	0.51	1.23	0.64	.036	.039	
51	0.19	0.34	2.21	0.43	−.007	.142	AA
52	0.08	0.44	2.85	0.60	−.033	.020	
53	0.21	0.19	3.63	0.22	−.012	.022	

	CTT		IRT		DIF		
	P	Point- biserial Corr.	b	a	Sex beta-uni	Ethnicity beta-uni	Group favored
54	0.09	0.44	2.59	0.61	−.108	.089	W, AA
55	0.25	0.40	1.54	0.48	−.113	.146	W, AA
56	0.24	0.32	1.96	0.37	−.041	.074	AA
57	0.25	0.27	2.66	0.26	.135	−.065	M, C
58	0.19	0.43	1.81	0.55	.135	−.160	M, C
59	0.14	0.28	3.36	0.34	.020	−.033	
60	0.18	0.49	1.66	0.66	.111	−.076	M, C
61	0.04	0.36	4.03	0.53	.000	.002	

M men, *W* women, *AA* African American, *C* Caucasian

The ranges of the difficulty values on the scale level were .07 to .55 (*M* = .31, *SD* = .12) for the Magical Ideation Scale, .04 to .42 (*M* = .17, *SD* = .11) for the Perceptual Aberration Scale, .04 to .67 (*M* = .21, *SD* = .14) for the Revised Social Anhedonia Scale, and .04 to .79 (*M* = .21, *SD* = .15) for the Physical Anhedonia Scale. The Perceptual Aberration Scale, the Revised Social Anhedonia Scale, and the Physical Anhedonia Scale include items that measure a broader range of the trait spectrum. This is true to a lesser degree for the Perceptual Aberration Scale, which seems to cover a smaller range.

Item Response Theory

For all four scales, overall results provided evidence for unidimensionality. Conclusions were based on scree plots, eigenvalues, fit indices, and factor loadings. The best fitting IRT model for the scales was the two-parameter IRT model. The guessing parameter was not needed because people low on the traits did not endorse items in a way that indicated schizotypy. Additional evidence for the two-parameter model resulted from the fit for people and fit for items with ZFIT—the least misfit applied for the two-parameter model.

Tables 2, 3, 4, and 5 display the psychometric item properties estimated in the IRT framework. They include discrimination parameters (a-values) and the difficulty parameters (b-values). Items with the highest discrimination values differentiate between people low and high on traits best. The ranges (means and SDs) of the discrimination values on the scale level were .12 to .95 (*M* = .63, *SD* = .21) for the Magical Ideation Scale, .21 to 1.66 (*M* = .96, *SD* = .36) for the Perceptual Aberration Scale, .17 to 1.15 (*M* = .67, *SD* = .27) for the Revised Social Anhedonia Scale, and .05 to 1.39 (*M* = .53, *SD* = .25) for the Physical Anhedonia Scale.

The discrimination potential within the IRT framework needs to be interpreted in combination with the difficulty parameters of the scales. The difficulty parameters' ranges (means and SDs) were .27 to 6.10 (*M* = 1.12, *SD* = 1.12) for the Magical Ideation Scale, .30 to 2.32 (*M* = 1.62, *SD* = .50) for the Perceptual Aberration Scale, -1.10 to 4.87 (*M* = 1.70, *SD* = .98) for the Revised Social Anhedonia Scale, and −14.82 to 6.21 (*M* = 1.89, *SD* = 2.52) for the Physical

Anhedonia Scale. All four scales are most discriminant at the higher end of the trait continuum. At the same time, in combination with the somewhat lower discrimination values, it can be concluded that a wider range of the higher trait continuum is covered for the Magical Ideation Scale, the Revised Social Anhedonia Scale, and especially for the Physical Anhedonia Scale. The Perceptual Aberration Scale, on the other hand, has the highest discrimination potential within a smaller range of the upper part of the trait continuum.

Information functions and standard-error functions provide additional insight into where on the trait scale and over which trait level range each scale currently provides most of the information. The test information maxima are 1.0 for the Magical Ideation Scale, 1.8 for the Perceptual Aberration Scale, 1.8 for the Revised Social Anhedonia Scale, and 2.0 for the Physical Anhedonia Scale. The information is highest around these maxima and hence the standard error is lowest in these trait ranges. At a trait level of about .5, the information is low and the standard error increases considerably. Consistent with the conclusions based on the discrimination parameters and difficulty parameters, the Perceptual Aberration Scale, the Revised Social Anhedonia Scale, and the Physical Anhedonia Scale measure most precisely at the high ends of the trait continuum, and the Magical Ideation Scale does so to a lesser degree.

Differential Item Functioning

DIF was performed with SIBTEST for sex and ethnicity (Stout and Roussos *1999*). Statistical results (beta-uni as effect sizes) are reported for each scale's items in Tables 2, 3, 4, and 5; Table 6 summarizes the DIF effects at the scale level. The results indicate that per scale, the following percentages of items display DIF (see Table 6): 47% of the Magical Ideation Scale, 23% of the Perceptual Aberration Scale, 48% of the Revised Social Anhedonia Scale, and 60% of the Physical Anhedonia Scale. The scales were differentiated by the percentage of items that fell into the category of moderate DIF (B-level) and high DIF (C-level). Whereas the C-level items for the Magical Ideation Scale and Perceptual Aberration Scale were small—7% and 6% respectively—the Revised Social Anhedonia Scale and the Physical Anhedonia Scale had large percentages of C-level items, 28% and 48%, respectively. Tables 2, 3, 4, and 5 provide information about which items favor which group—men vs. women and African Americans vs. Caucasians.

Table 6 Number and percent of items displaying differential item functioning (DIF) overall and number and percent of items displaying moderate DIF (B-level) and large DIF (C-level)

	Overall	Overall%	B-level	B-level%	C-level	C-level%
Magical Ideation	14	47	12	40	2	7
Perceptual Aberration	8	23	6	17	2	6
Social Anhedonia	19	48	8	20	11	28
Physical Anhedonia	36	60	7	12	29	48

Discussion

This study examined the psychometric properties of the WSS according to CTT, IRT, and DIF. The different measurement frameworks provide results that are consistent with each other. According to CTT and IRT, the Perceptual Aberration Scale does the best job of differentiating between people low and high on the trait, whereas the Physical Anhedonia Scale's potential to differentiate is somewhat lower. The Revised Social Anhedonia Scale and the Magical Ideation Scale are intermediate in that respect. This pattern is also reflected in the reliability coefficients: internal consistency is highest for the Perceptual Aberration Scale and somewhat lower for the other scales. The Physical Anhedonia Scale has acceptable internal consistency, but it has the most items of the four scales. CTT and IRT results also confirm that the Wisconsin Schizotypy Scales concentrate their measurement on the higher end of the trait continuum, which was intended by the original scale developers.

When looking at the DIF results, there is overlapping information with CTT and IRT. Items that function differently for groups often also have lower discrimination values. In general, DIF items overall, but especially the ones displaying high DIF, are predominantly in the Revised Social Anhedonia Scale and the Physical Anhedonia Scale. This is in line with previous research that found differences for sex and ethnicity for these scales (Chmielewski et al.*1995*; Kwapil et al. *2002*). For each scale, there are items that "favor" men and women, as well as African Americans and Caucasians. Based on these results, we could conclude that the differences for individual items for sex and ethnicity are not real differences in the traits intended to be measured. The consequence of finding DIF for these scales is that the score validity might differ for the subgroups. According to the standards for educational and psychological measurement (AERA, APA, NCME *1999*), it would be appropriate to provide evidence for score validity for the different subgroups. In the meantime, we recommend that researchers use different subgroup norms.

This study extends earlier Rasch analyses of the scales by Graves and Weinstein (*2004*) in several respects. First, our results are based on 6,137 subjects (versus 90 in the Graves and Weinstein study) to increase the stability of the estimates. Second, a two-parameter model was applied versus a Rasch model to accommodate items related to the construct to differing degrees. Third, this study investigated DIF for sex and ethnicity, which Graves and Weinstein suggested future research ought to include. In addition, this study extends a recent generalizability analysis of the Wisconsin scales (Winterstein et al. *2010*). That analysis found several substantial item-by-person interactions, which suggested that some items may mean different things to different subgroups. The present analyses build upon those findings by formally testing for DIF. The results indicated significant DIF in each scale, consistent with the prior analyses.

Several limitations of the present work should be noted. Although the sample size was large, the sample consisted of young adults enrolled in a university. College students are within the window of risk for many schizophrenia-spectrum disorders, but as a whole they may have characteristics and protective factors that limit the generalizability of the findings to the population of young adults more broadly. Similarly, Caucasians and African-Americans were the only racial groups represented in this sample. A worthwhile task for future work would be to

seek large samples that include other age ranges and racial and ethnic groups and that include people who are at risk for or diagnosed with schizophrenia and related disorders. Finally, it would be worthwhile for future work to evaluate several schizotypy scales, not just the WSS. The WSS are among the most widely used measures of schizotypy, but there are many popular scales (e.g., Schizotypal Personality Questionnaire; Raine *1991*; Oxford-Liverpool Inventory of Feelings and Experiences; Mason et al. *1995*). A comparative psychometric evaluation, particularly with regards to DIF, would illuminate the strengths and weaknesses of the WSS.

The present results offer useful information for applied users of the Wisconsin Schizotypy Scales. First, the scales effectively assess the trait range intended by the scale developers—the test information functions peak at the traits' high end. As a consequence, people with high scores are measured more reliably, and people with low scores are measured less reliably. Second, the scales fared well overall, but it is clear that DIF is a serious issue. Some of the scales had substantial percentages of high-DIF items; almost half of the Physical Anhedonia Scale's items, for example, were high-DIF items. The widespread DIF in these scales indicates a need for caution when interpreting sex and ethnic differences that have been found in the literature: such differences may not reflect true trait differences. Taken together, the findings suggest that further psychometric development of the WSS is warranted.

References:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington: AERA.

Barrantes-Vidal, N., Lewandowski, K. E., & Kwapil, T. R. (2010). Psychopathology, social adjustment and personality correlates of schizotypy clusters in a large nonclinical sample. Schizophrenia Research, 122, 219–225.

Barrantes-Vidal, N., Ros-Morente, A., & Kwapil, T. R. (2009). An examination of neuroticism as a moderating factor in the association of positive and negative schizotypy with psychopathology in a nonclinical sample. Schizophrenia Research, 115, 303–309.

Chapman, L. J., Chapman, J. P., Kwapil, T. R., Eckblad, M., & Zinser, M. C. (1994). Putatively psychosis-prone subjects 10 years later. Journal of Abnormal Psychology, 103, 171–183.

Chapman, L. J., Chapman, J. P., & Raulin, M. L. (1976). Scales for physical and social Anhedonia. Journal of Abnormal Psychology, 85, 374–382.

Chapman, L. J., Chapman, J. P., & Raulin, M. L. (1978). Body image aberration in schizophrenia. Journal of Abnormal Psychology, 87, 399–407.

Chmielewski, P. M., Fernandes, L. O. L., Yee, C. M., & Miller, G. A. (1995). Ethnicity and gender in scales of psychosis proneness and mood disorders. Journal of Abnormal Psychology, 104, 464–470.

Claridge, G., McCreery, C., Mason, O., Bentall, R., Boyle, G., Slade, P., et al. (1996). The factor structure of “schizotypal” traits: a large replication study. British Journal of Clinical Psychology, 35, 103–115.

DeVellis, R. F. (2012). Scale development: Theory and applications (3rd ed.). Los Angeles: Sage.

Eckblad, M., & Chapman, L. J. (1983). Magical Ideation as an indicator of schizotypy. Journal of Consulting and Clinical Psychology, 51, 215–225.

Eckblad, M., Chapman, L. J., Chapman, J. P., & Mishlove, M. (1982). The revised social Anhedonia scale. Unpublished test (copies available from T. R. Kwapil, Department of Psychology, University of North Carolina at Greensboro, P.O. Box 26170 Greensboro, NC 27402–6170).

Gooding, D. C., Tallent, K. A., & Matts, C. W. (2005). Clinical status of at-risk individuals 5 years later: further validation of the psychometric high-risk strategy. Journal of Abnormal Psychology, 114, 170–175.

Gooding, D. C., Tallent, K. A., & Matts, C. W. (2007). Rates of avoidant, schizotypal, schizoid and paranoid personality disorders in psychometric high-risk groups at 5-year follow-up. Schizophrenia Research, 94, 373–374.

Graves, R. E., & Weinstein, S. (2004). A rasch analysis of three of the Wisconsin scales of psychosis proneness: measurement of schizotypy. Journal of Applied Measurement, 5, 160–171.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), Current topics in clinical and community psychology (Vol. 2, pp. 61–96). San Diego: Academic.

Kaczorowski, J. A., Barrantes-Vidal, N., & Kwapil, T. R. (2009). Neurological soft signs in psychometrically identified schizotypy. Schizophrenia Research, 115, 293–302.

Kwapil, T. R. (1998). Social Anhedonia as a predictor of the development of schizophrenia-spectrum disorders. Journal of Abnormal Psychology, 107, 558–565.

Kwapil, T. R., Barrantes-Vidal, N., & Silvia, P. J. (2008). The dimensional structure of the Wisconsin Schizotypy Scales: factor identification and construct validity. Schizophrenia Bulletin, 34, 444–457.

Kwapil, T. R., Crump, R. A., & Pickup, D. R. (2002). Assessment of psychosis proneness in African-American college students. Journal of Clinical Psychology, 58, 1601–1614.

Mason, O., Claridge, G., & Jackson, M. (1995). New scales for the assessment of schizotypy. Personality and Individual Differences, 18, 7–13.

Meehl, P. E. (1990). Toward an integrated theory of schizotaxia, schizotypy, and schizophrenia. Journal of Personality Disorders, 4, 1–99.

Raine, A. (1991). The SPQ: a scale for the assessment of schizotypal personality based on DSM-III-R criteria. Schizophrenia Bulletin, 17, 555–564.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14, 127–137.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performances. Journal of Educational Measurement, 33, 215–230.

Silvia, P. J., & Kwapil, T. R. (2010). Aberrant asociality: How individual differences in social Anhedonia illuminate the need to belong. Journal of Personality. doi:10.1111/j.1467-6494.2010.00702.x

Stout, W., & Roussos, L. (1999). Dimensionality-based DIF/DBF package [Computer program]. Champaign-Urbana: William Stout Institute for Measurement, University of Illinois.

Winterstein, B. P., Willse, J. T., Kwapil, T. R., & Silvia, P. J. (2010). Assessment of score dependability of the Wisconsin Schizotypy Scales using generalizability analysis. Journal of Psychopathology and Behavioral Assessment, 32, 575–585.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models. Chicago: Scientific Software International.