Assessment of Score Dependability of the Wisconsin Schizotypy Scales Using Generalizability Analysis

By: Beate P. Winterstein, John T. Willse, Thomas R. Kwapil and Paul J. Silvia

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 & Behavioral Assessment, 32,* 575-585. http://dx.doi.org/10.1007/s10862-010-9181-x

Abstract:

To investigate the reliability of the Wisconsin Schizotypy Scales, this study applied generalizability analysis with two college student samples who completed the scales at two time points. The results indicated that the Revised Social Anhedonia Scale had acceptable levels of score dependability, but that the score dependability for the other scales (the Physical Anhedonia Scale, the Perceptual Aberration Scale, and the Magical Ideation Scale) was below an acceptable level of .80. The patterns of variance components suggested that the scales' items need improvement. Researchers can use the included tables to choose the number of items and occasions needed to get dependable score interpretations. This research was presented at the 2007 meeting of the Midwestern Psychological Association, Chicago, IL.

Keywords: schizotypy | schizophrenia | wisconsin schizotypy scales | generalizability theory | dependability | reliability | psychology

Article:

This study investigated score dependability of the Wisconsin Schizotypy Scales: the Revised Social Anhedonia Scale (Chapman et al. 1976; Eckblad et al. 1982), the Physical Anhedonia Scale (Chapman et al. 1976), the Perceptual Aberration Scale (Chapman et al. 1978), and the Magical Ideation Scale (Eckblad and Chapman 1983) in the framework of Generalizability Theory (Cronbach et al. 1972). Generalizability Theory (G-Theory), an extension of Classical Test Theory (CTT), considers many sources of error at a time (Brennan 2001; Shavelson and Webb 1991). As a consequence, score reliability—or score dependability, as it is called in Generalizability Theory—accounts for all defined error terms. Understanding error is central to enhancing assessment quality, and G-Theory offers unique insights into the many sources of error.

Schizotypy

Current diagnostic formulations define schizophrenia as a categorical disorder. However, schizophrenia is often construed as the most extreme manifestation of a continuum of clinical and subclinical impairment referred to as schizotypy (Meehl *1989*). Therefore, the schizotypy continuum subsumes schizophrenia and spectrum disorders (e.g., schizotypal personality disorder) as well as constructs such as the prodrome (McGlashan et al. *2001*). Most schizotypic people will never develop psychotic or spectrum disorders; however, they may experience mild or transient signs of schizophrenic impairment, including magical thinking, odd perceptual experiences, neurocognitive impairment, and social impairment. The reliable identification of schizotypic people should enhance our understanding of the etiology of schizophrenia and related conditions, and it is essential for the development and application of prophylactic treatment interventions.

The Chapmans and their collaborators developed a series of self-report, true-false questionnaires that were intended to measure symptoms and traits reported to be characteristic of schizotypy. These included the Magical Ideation (Eckblad and Chapman *1983*), Perceptual Aberration (Chapman et al. *1978*), Physical Anhedonia (Chapman et al. *1976*), and Revised Social Anhedonia (Eckblad et al. *1982*) Scales. The content of the measures followed largely from Meehl's (*1964*) checklist of schizotypic signs. The scales were developed following Jackson's (*1970*) recommendations for the construction of personality measures resulting in internally consistent questionnaires. Candidate items were administered to large samples to screen them for discriminant validity, gender bias, acquiescence, and social desirability. Furthermore, given the expected low base rate of schizotypy in the general population and the authors' interest in identifying participants with high levels of schizotypic traits, items were selected that had low endorsement frequencies. This produced positively skewed scales that were maximally discriminant at the high end of the scale (Chapman et al. *1995*).

Coefficient alpha reliabilities for the scales range from .79 to .90, and test-retest reliability over six weeks ranged from .75 to .82 (Chapman et al. *1982*; Kwapil et al. *2008*; Mishlove and Chapman *1985*). The Wisconsin Schizotypy Scales have been widely used in cross-sectional and longitudinal studies of schizotypy. High scorers on the scales exhibit schizophrenic-like cognitive, emotional, and social impairment, and are at elevated risk for developing schizophrenia-spectrum disorders and other disorders (e.g., Chapman et al. *1994*; Chmielewski et al. *1995*; Edell *1995*; Kwapil *1998*).

Generalizability Theory

G-Theory (Brennan *2001*; Cronbach et al. *1972*; Shavelson and Webb *1991*) is appealing for research on reliability because it considers many sources of error at once. Traditional methods of estimating reliability can handle only one dimension of variability, known as a facet. For example, researchers interested in reliability across items (an item facet) can use Cronbach's alpha, researchers interested in reliability across raters (a rater facet) can use measures of

agreement (e.g., kappa), and researchers interested in reliability across time (an occasion facet) can use test-retest reliability.

But many assessment contexts have more than one facet. For instance, a set of items may be administered twice. Alpha will estimate item reliability at each time point, and test-retest reliability will estimate reliability across time, but neither will estimate a single, holistic summary of the reliability of the scores yielded by the two-facet item-by-occasion design. Similarly, a study might ask participants to complete three tasks and have four raters score each task. In short, a holistic estimate of reliability is needed that can take into account both facets (tasks and raters) but also the interactions between the participants' trait levels, the tasks, and the raters. G-theory thus provides a generalizability coefficient that, like Cronbach's alpha, ranges from 0 to 1, with higher levels indicating more dependable scores. Values of .80 and higher are usually viewed as acceptable.

Cronbach et al.'s (*1972*) G-theory is thus a logical extension of Cronbach's earlier work on item reliability. G-theory extends studies of reliability to research involving more than one facet, and it takes into account all defined sources of error for a given design. The value of doing so is twofold. First, understanding error is central to improving assessment quality. By illuminating the major sources of error in a research design, G-theory can point researchers to the facets that need attention. If raters contribute a lot of variance in scores, for example, then researchers can either add raters or retrain the raters; if items contribute a lot of variance, then researchers can add or revise the items.

Second, G-theory enables forward-looking estimates of what score dependability would be for different research designs. Once the contributions of the different sources of error are known, researchers can estimate how alternate designs would affect dependability. Just as the Spearman-Brown formula (Brown *1910*; Spearman *1910*) allows researchers to forecast what alpha reliability would be if items were added, G-theory allows researchers to forecast what the holistic dependability score would be if the various facets were at different levels. This enables researchers to plan research that optimizes the likelihood of obtaining dependable scores. For example, a recent study of creativity assessment found that the number of creativity tasks was much less important than the number of raters, and that returns in dependability diminished quickly after 3 or 4 raters (Silvia et al. *2008*). Future researchers can thus devote their resources to recruiting and training raters instead of administering more tasks.

To apply G-Theory, researchers plan two analyses, known as "studies": a generalizability study (G-Study) and a decision study (D-Study). For the G-Study, a universe of admissible observations (which consists of the object of measurement, typically examinees, and the measurement error facets) is defined and the variance components are estimated. For the D-Study, a universe of admissible generalizations (which represents the measurement conditions based on the object of measurement and the measurement facets a researcher is willing to generalize over) is defined and the variance components associated with the universe of admissible generalizations are estimated. Examples of measurement facets in G-Studies and D-Studies are items, occasions, raters, interviewers, and scoring methods. Facets are defined as similar measurement conditions and are considered interchangeable. For example, if researchers

study examinees' expression of social anhedonia, then they are interested in someone's trait level but not what someone's score is on a particular set of items on a particular occasion. Hence, all observations resulting from the facets in a defined universe of admissible observations are considered interchangeable, which results in the researcher's willingness to generalize from an observed score to the expected mean score over an infinite number of conditions for each facet.

Furthermore, researchers can estimate dependability for two measurement approaches: norm-referenced testing and criterion-referenced testing. If one's interest lies in ranking people (relative decision), then the Generalizability coefficient (G-coefficient) informs about how dependable a score is. If one's interest lies in the absolute standing to a criterion (absolute decision), then the Phi-coefficient (Φ coefficient) reflects the score dependability. Phi-coefficients are typically lower than G-coefficients because they consider the main error effects in addition to the interaction effects that are used for G-coefficients.

In short, G-Theory answers two questions: How dependable is a measurement instrument with the current design, and how dependable would it be for alternative designs? To answer these questions, G-Theory follows a descriptive approach by reporting estimated variance components and estimated dependability coefficients. Consequently, it makes weak assumptions. It assumes that all score effects (main and interaction) are uncorrelated, but not that the effects are independent. Most important, it is not necessary to assume normality because no inferential generalizations are made (i.e., significance tests). For more about G-Theory, see Shavelson and Webb (*1991*) for an overview, Brennan (*2001*) and Cronbach et al. (*1972*) for more technical elaborations, and other studies in clinical and counseling research (e.g., Hoyt and Melby *1999*; Nugent *2006*; Seeger Halvorsen et al. *2006*).

The Present Research

We applied G-Theory to investigate the score dependability of the Revised Social Anhedonia, Physical Anhedonia, Perceptual Aberration, and Magical Ideation Scales in two independent samples of undergraduate students collected at different time points—sample 1 (primary sample) and sample 2 (replication sample). The second sample was used to investigate the stability of the statistics (variance components and dependability coefficients). G-Studies and D-Studies were performed to estimate the variance components and dependability coefficients in the univariate two-facet person-by-occasion-by-item design (p × o × i design). Based on the G-Theory approach, the following research questions were answered: How dependable are scores from the Wisconsin Schizotypy Scales? What are the major sources of error? What alternative designs (different number of occasions or items) would be required for the scales to reach desired levels of dependability?

For this study, we defined items and occasions as two facets of measurement error with interchangeable measurement conditions: items might differ in the level of the construct they measure, and occasions might differ in situational (such as time of the day, room temperature) as well as individual (such as general mood, alertness level of the examinee) variables. The design is balanced (the same items are applied at both occasions) and the facets are crossed (every examinee participates at both occasions and replies to all items).

Method

Participants

Undergraduate students enrolled in General Psychology classes at the University of North Carolina at Greensboro participated. Sample 1 ($n = 160$) was 66% female and 34% male; regarding ethnicity, it was 68% Caucasian, 22% African American, 2% Asian, 3% Hispanic, 1% Native American, 1% other, and 3% participants who did not indicate their ethnicity. Sample 2 was 73% female and 27% male; regarding ethnicity, it was 74% Caucasian, 17% African-American, 6% Asian, 1% other, and 2% participants who did not indicate their ethnicity.

Materials and Procedures

Participants in both samples were administered the Revised Social Anhedonia, Physical Anhedonia, Perceptual Aberration, and Magical Ideation Scales. All four measures have true-false items, and the score for each scale is the sum of responses in the deviant direction. The Revised Social Anhedonia Scale has 40 items that assess lack of social interest and pleasure, the Physical Anhedonia Scale has 61 items that assess lack of sensory and aesthetic pleasure, the Perceptual Aberration Scale has 35 items that assess schizophrenic-like perceptual and bodily distortions, and the Magical Ideation Scale has 30 items that assess belief in implausible forms of causation.

The items on the schizotypy scales were intermixed with a 13-item measure of infrequent responding (Chapman and Chapman *1983*). The infrequency scale was included to screen out participants who responded in a random or “fake-bad” manner. Consistent with the recommendations of Chapman and Chapman (*1983*), participants who endorsed more than two infrequency items were dropped from further study and thus are not included in the present samples or analyses. Participants completed these measures (along with measures not used in this study) for course credit as part of group mass-screening sessions that lasted 1.5 to 2 h; the two time points were separated by 8 to 12 weeks.

Statistical Method

The present study employed GENOVA (Crick and Brennan *1983*) for the univariate two-facet person-by-occasion-by-item design ($p \times o \times i$). We performed two studies for each scale, a G-Study and a D-Study. In the G-Studies, we estimated the variance components associated with the object of measurement (the examinees), the measurement facets (occasions and items), and their interactions for all scales. The G-Studies also provided the dependability for the current design for all four scales. In the D-Studies, we estimated variance components and dependability coefficients associated with different designs for all scales. We used a dependability threshold of .80 as a standard for acceptable reliability (DeVellis *2003*).

Results

Descriptive Statistics and Traditional Reliability Coefficients

Table 1 displays the descriptive statistics for all five scales. All scales produce distributions that are positively skewed; the Perceptual Aberration Scale and the Revised Social Anhedonia Scale show the most skew. Concerning traditional reliability coefficients, Table 2 shows Cronbach's alpha and test-retest reliability for all four scales. Most of the Cronbach's alphas indicate acceptable levels of score reliability—they range from .79 to .91. The test-retest reliability coefficients are at an acceptable level for the Revised Social Anhedonia Scale (.81 for sample 1 and sample 2) and the Physical Anhedonia Scale (.81 for sample 1 and sample 2), but below the threshold for the Magical Ideation Scale (.73 for sample 1 and .79 for sample 2) and the Perceptual Aberration Scale (.63 for sample 1 and .76 for sample 2).

Table 1 Descriptive statistics for the Revised Social Anhedonia Scale, the Physical Anhedonia Scale, the Perceptual Aberration Scale, and the Magical Ideation Scale (Sample 1 & 2)

		N		Mean		SD		Var		Skewness		Kurtosis	
		S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
Revised Social Anhedonia	O1	160	100	8.21	7.96	5.80	5.45	33.69	29.65	1.62	1.57	4.61	3.23
	O2			7.25	7.68	5.36	5.69	28.74	32.32	1.33	1.22	2.52	1.44
Physical Anhedonia	O1	148	94	13.37	13.98	7.58	7.21	57.49	51.91	0.94	0.56	0.68	−0.42
	O2			11.96	12.26	6.99	7.32	48.91	53.55	0.86	1.21	0.89	1.55
Magical Ideation	O1	159	101	9.19	10.16	5.05	5.72	25.54	32.71	0.57	0.64	−0.35	−0.23
	O2			8.72	9.10	5.63	5.98	31.71	35.77	0.65	0.81	−0.24	0.22
Perceptual Aberration	O1	156	102	5.31	6.48	4.65	5.57	21.64	31.02	1.49	1.46	2.85	2.80
	O2			4.33	4.66	5.31	5.60	28.20	31.36	2.52	2.37	7.97	6.54

SD standard deviation, *Var* variance, *O1* occasion 1, *O2* occasion 2, *S1* Sample 1, *S2* Sample 2

Table 2 Traditional reliability coefficients—Cronbach's alpha and test-retest reliability—for the Revised Social Anhedonia Scale, the Physical Anhedonia Scale, the Perceptual Aberration Scale, and the Magical Ideation Scale (Sample 1 & 2)

	Cronbach's alpha				Test-retest reliability	
	Sample 1		Sample 2		Sample 1	Sample 2
	O1	O2	O1	O2		
Revised Social Anhedonia	.85	.84	.83	.85	.81	.81
Physical Anhedonia	.86	.85	.83	.86	.81	.81
Magical Ideation	.79	.85	.84	.86	.73	.79
Perceptual Aberration	.83	.90	.87	.91	.63	.76

O1 occasion 1, *O2* occasion 2

The Revised Social Anhedonia Scale

Table 3 provides estimated G-study results based on the current design of sample 1, in which 160 students filled out the Revised Social Anhedonia Scale with 40 items at two different occasions. Only a small source of variation (9%) in peoples' scale scores was due to differences among people in expressing social anhedonia. This implies that effects other than the person's level of social anhedonia influences the score on the Revised Social Anhedonia Scale, namely items (12%) and person-by-item interaction (28%). This result suggests that items show some inconsistency by possibly measuring different levels of social anhedonia (the item variance). People respond inconsistently across items (the person-by-item interaction), which means that different people might understand and react to the same item in different ways despite having the same total score on the scale. Little variance is introduced by occasions alone (0%), the person-by-occasion interaction (1%), and the occasion-by-item interaction (0%).

Table 3 Revised Social Anhedonia Scale: estimated variance components, standard error (SE), and percentage of variance accounted for by effects (percent) for p × o × i design (Sample 1 & 2)

Effects	Variance		SE		Percent	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
p	0.0146	0.0144	0.0020	0.0024	9.33	9.14
o	0.0003	0.0	0.0002	0.0	0.17	0.0
i	0.0185	0.0204	0.0042	0.0047	11.87	12.93
p × o	0.0019	0.0017	0.0004	0.0005	1.21	1.10
p × i	0.0432	0.0465	0.0016	0.0021	27.67	29.47
o × i	0.0003	0.0001	0.0002	0.0002	0.16	0.03
p × o × i, e	0.0774	0.0747	0.0014	0.0017	49.59	47.34

p person effect, *o* occasion effect, *i* item effect, *p* × *o* person-by-occasion interaction effect, *p* × *i* person-by-item interaction effect, *o* × *i* occasion-by-item interaction effect, *p* × *o* × *i*, *e* person-by-occasion-by-item interaction plus error effect

The largest source of variation in peoples' scale scores (50%) is accounted for by the 3-way interaction confounded with random error. In G-studies, this component has no clear interpretation because it confounds the interaction with random error (Brennan *2001*). There is thus no way to determine if this large variance component is due to the 3-way interaction between people-by-items-by-occasion, if there is some other currently unidentified facet that needs to be investigated in the future, or if random error is responsible. In conclusion, most error is associated with items, and the error for occasions is negligible.

Table 4 displays the G-coefficients and Phi-Coefficients that were estimated for the current design of sample 1 (2 occasions and 40 items) and alternative designs with varied numbers of occasions and items. We estimated a G-coefficient of .83 and a Phi-Coefficient of .80 for the current design, which can be considered acceptable based on a standard of .80. Hence, the current design with 2 occasions and 40 items is efficient. For an alternative design with 1 occasion and 40 items, the G-Coefficient is estimated to be .75 and the Phi-Coefficient .72. In the

case of 3 occasions and 40 items, we would expect a G-Coefficient of .86 and a Phi-Coefficient of .83. Hence, the increase from 0.75 (one occasion) to 0.83 (two occasions) is noticeable with an increase of .08 and might justify the extra effort of asking participants to fill out the scale a second time. Diminishing returns then appear because the increase from 2 to 3 occasions is only .03. Hence, adding additional occasions beyond 2 wouldn't contribute much to a higher dependability of scores.

Table 4 Revised Social Anhedonia Scale: G-coefficients and Φ-coefficients for different numbers of occasions and items for p × O × I design (Sample 1 & 2)

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
1	25	0.68	0.69	0.65	0.66
	30	0.71	0.71	0.68	0.69
	35	0.73	0.74	0.70	0.71
	40	0.75	0.75	0.72	0.73
	45	0.76	0.77	0.74	0.75
	50	0.77	0.78	0.75	0.76
2	25	0.78	0.77	0.74	0.74
	30	0.80	0.80	0.77	0.77
	35	0.82	0.82	0.79	0.79
	40	0.83	0.83	0.80	0.81
	45	0.84	0.84	0.81	0.82
	50	0.85	0.85	0.83	0.83
3	25	0.81	0.81	0.78	0.77
	30	0.83	0.83	0.80	0.80
	35	0.85	0.85	0.82	0.82
	40	0.86	0.86	0.83	0.83
	45	0.87	0.87	0.85	0.85
	50	0.88	0.88	0.86	0.86

At first glance, the most efficient approach to enhance dependability would be to add items because the person-by-item variance (28%) and item variance (12%) are proportionally large sources of error. Typically, when large amounts of error are associated with a facet, it can be reduced by adding conditions for the facet. But as can be seen in Table 4, adding items doesn't provide much benefit after the design includes about 35 items, especially when implementing the scale two and three times.

In sample 2 (replication sample), we found similar results for the variance components and their standard errors, the percentages explained by the effects, and dependability coefficients.

The Physical Anhedonia Scale

As can be seen in Table 5 for sample 1, only a small amount of variance (6%) is accounted for by participants' real differences on the construct. Occasions (0%) and their interaction with the object of measurement (1%) and with items (0%) explain only trivial percentages of variance. Items (14%) and their interaction with persons (32%) introduce a substantial amount of error. The confounded 3-way interaction and random error account for 46% of the score variance.

Table 5 Physical Anhedonia Scale: estimated variance components, standard error (SE), and percentage of variance accounted for by effects (percent) for p × o × i design (Sample 1 & 2)

Effects	Variance		SE		Percent	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
p	0.0106	0.0106	0.0015	0.0019	6.44	6.24
o	0.0002	0.0004	0.0002	0.0003	0.15	0.22
i	0.0230	0.0236	0.0042	0.0044	13.91	13.93
p × o	0.0015	0.0014	0.0003	0.0004	0.93	0.83
p × i	0.0533	0.0570	0.0015	0.0019	32.26	33.63
o × i	0.0	0.0002	0.0001	0.0002	0.01	0.13
p × o × i, e	0.0765	0.0763	0.0012	0.0014	46.30	45.01

p person effect, *o* occasion effect, *i* item effect, *p* × *o* person-by-occasion interaction effect, *p* × *i* person-by-item interaction effect, *o* × *i* occasion-by-item interaction effect, *p* × *o* × *i*, *e* person-by-occasion-by-item interaction plus error effect

In terms of dependability coefficients for sample 1, Table 6 shows that for 2 occasions and 61 items each the estimated G-coefficient is .80 and the estimated Phi-coefficient is .77. See Table 6 for estimated dependability coefficients for alternative designs. Based on the results, we can conclude that dependability is gained by applying the scale with 61 items at two occasions (estimated dependability goes up from .73 to .80 for relative interpretations and from .70 to .77 for absolute interpretations). But there are diminishing returns when adding a third occasion (a gain of only .02). Concerning items, the conclusions made for the Revised Social Anhedonia Scale also apply here: although the variances accounted for by items (14%) and person-by-item interaction (32%) are substantial, there is only gain in score dependability up to a certain number of items (around 35 to 45, depending on the number of occasions). This implies that the items' quality, not quantity, should be addressed.

Table 6 Physical Anhedonia Scale: G-coefficients and Φ-coefficients for different numbers of occasions and items for p × O × I design (Sample 1 & 2)

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
1	25	0.61	0.61	0.57	0.57
	30	0.64	0.64	0.61	0.60
	35	0.67	0.67	0.63	0.63
	40	0.69	0.69	0.65	0.65
	45	0.71	0.71	0.67	0.67
	50	0.72	0.72	0.69	0.68
	55	0.73	0.73	0.70	0.70
	61	0.74	0.75	0.71	0.71
2	25	0.71	0.70	0.66	0.65
	30	0.74	0.73	0.69	0.69
	35	0.76	0.76	0.72	0.71
	40	0.78	0.77	0.74	0.73
	45	0.79	0.79	0.76	0.75
	50	0.80	0.80	0.77	0.76
	55	0.81	0.81	0.78	0.78
	61	0.82	0.82	0.79	0.79
3	25	0.74	0.74	0.69	0.69
	30	0.77	0.77	0.73	0.72
	35	0.79	0.79	0.75	0.74
	40	0.81	0.81	0.77	0.77
	45	0.82	0.82	0.79	0.78
	50	0.84	0.83	0.80	0.80
	55	0.85	0.84	0.81	0.81
	61	0.85	0.85	0.82	0.82

As was the case for the Revised Social Anhedonia Scale, the variance component estimates as well as the dependability coefficient estimates in sample 2 resemble those in sample 1 closely. Hence, the stability over samples with similar characteristics was high.

The Perceptual Aberration Scale

As can be seen in Table 7 for sample 1, the variance accounted for by people's level of perceptual aberration is small (10%). Other effects contributed to the overall score variance: items (7%) and person-by-item interaction (17%). With 60%, the 3-way interaction confounded with random error is the biggest source of error. No variance is accounted for by occasions, by the occasion-by-item interaction, and only 5% by the person-by-occasion interaction. The

interpretation differs from the Revised Social Anhedonia Scale (see above) and the Magical Ideation Scale (see below). Since only small amounts of variance are explained by occasions and items and their respective interactions with the object of measurement and other facets, there will not be much room for improving the measurement procedure by adding items or occasions. For the current design with 35 items at each of two occasions, we estimated a G-Coefficient of .73 and a Phi-Coefficient of .71. Hence, the score dependability for the current design is below the acceptable level of .80. To reach an acceptable score dependability of above .80, the design would require 40 items each at three occasions for relative decisions and 50 items for absolute decisions (Table 8).

Table 7 Perceptual Aberration Scale: estimated variance components, standard error (SE), and percentage of variance accounted for by effects (percent) for p × o × i design (Sample 1 & 2)

Effects	Variance		SE		Percent	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
p	0.0120	0.0186	0.0019	0.0032	10.08	13.80
o	0.0003	0.0013	0.0003	0.0011	0.28	0.94
i	0.0085	0.0104	0.0021	0.0027	7.12	7.69
p × o	0.0057	0.0040	0.0009	0.0008	4.74	2.96
p × i	0.0205	0.0292	0.0013	0.0018	17.19	21.64
o × i	0.0003	0.0011	0.0002	0.0004	0.29	0.78
p × o × i, e	0.0720	0.0704	0.0014	0.0017	60.29	52.19

p person effect, *o* occasion effect, *i* item effect, *p* × *o* person-by-occasion interaction effect, *p* × *i* person-by-item interaction effect, *o* × *i* occasion-by-item interaction effect, *p* × *o* × *i*, *e* person-by-occasion-by-item interaction plus error effect

Table 8 Perceptual Aberration Scale: G-coefficients and Φ-coefficients for different numbers of occasions and items for p × O × I design (Sample 1 & 2)

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
1	25	0.56	0.70	0.55	0.66
	30	0.58	0.72	0.56	0.67
	35	0.59	0.73	0.58	0.69
	40	0.60	0.74	0.59	0.70
	45	0.61	0.75	0.59	0.71
	50	0.62	0.76	0.60	0.71
2	25	0.70	0.80	0.68	0.77
	30	0.72	0.82	0.70	0.78

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
	35	0.73	0.83	0.71	0.80
	40	0.74	0.84	0.72	0.80
	45	0.75	0.84	0.73	0.81
	50	0.75	0.85	0.74	0.82
3	25	0.77	0.84	0.74	0.81
	30	0.78	0.86	0.76	0.83
	35	0.79	0.87	0.77	0.84
	40	0.80	0.88	0.78	0.85
	45	0.81	0.88	0.79	0.85
	50	0.81	0.89	0.80	0.86

In comparison to the Revised Social Anhedonia Scale and the Physical Anhedonia Scale, the results from sample 1 differ to those in sample 2. The results look more favorable for score dependability in sample 2: for relative interpretations, the estimated G-coefficient is .83 and for absolute decisions the estimated phi-coefficient is .80. The variance attributable to real difference between people on perceptual aberration improved (from 10% in sample 1 to 14% in sample 2) while the confounded error (p × o × i, e) decreased (from 60% in sample 1 to 52% in sample 2). Hence, the stability of statistics is somewhat less favorable here.

The Magical Ideation Scale

In Table 9 for sample 1, the G-Study results are presented. Again, the variance accounted for by people's differences in expressing magical ideation is small (10%), and the items (6%) and the person-by-item interaction (30%) explains a good amount of the variance. The 3-way interaction confounded with random error represents the largest source of variation (51%). The person-by-occasion interaction (2%), the occasion alone (0%), and the occasion-by-item interaction (0%) introduce little variance.

Table 9 Magical Ideation Scale: estimated variance components, standard error (SE), and percentage of variance accounted for by effects (percent) for p × o × i design (Sample 1 & 2)

Effects	Variance		SE		Percent	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
p	0.0209	0.0277	0.0031	0.0048	9.95	12.66
o	0.0001	0.0005	0.0001	0.0005	0.03	0.24
i	0.0134	0.0147	0.0036	0.0041	6.37	6.71
p × o	0.0052	0.0046	0.0010	0.0011	2.49	2.10

Effects	Variance		SE		Percent	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
p × i	0.0627	0.0642	0.0027	0.0034	29.85	29.33
o × i	0.0003	0.0006	0.0002	0.0004	0.14	0.28
p × o × i, e	0.1075	0.1066	0.0022	0.0028	51.18	48.68

p person effect, *o* occasion effect, *i* item effect, *p* × *o* person-by-occasion interaction effect, *p* × *i* person-by-item interaction effect, *o* × *i* occasion-by-item interaction effect, *p* × *o* × *i*, *e* person-by-occasion-by-item interaction plus error effect

In terms of G-coefficients and Phi-Coefficients for sample 1, Table 10 provides guidelines. A G-coefficient of .76 and a Phi-Coefficient of .75 were estimated (which are below .80) for the current design with the same 30 items at 2 occasions. An acceptable level of about .80 can be reached by either having 45 items each for 2 occasions or 30 items each for 3 occasions. Hence, researchers need to decide if it's more beneficial to add items or to apply the scale three times. Both options might not be ideal for applications in research or practice.

Table 10 Magical Ideation Scale: G-coefficients and Φ-coefficients for different numbers of occasions and items for p × O × I design (Sample 1 & 2)

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
1	20	0.60	0.68	0.59	0.66
	25	0.63	0.71	0.62	0.69
	30	0.66	0.73	0.65	0.71
	35	0.67	0.75	0.66	0.73
	40	0.69	0.76	0.68	0.74
	45	0.70	0.77	0.69	0.75
	50	0.71	0.78	0.70	0.76
	55	0.72	0.78	0.71	0.77
2	20	0.71	0.77	0.70	0.75
	25	0.74	0.80	0.73	0.78
	30	0.76	0.82	0.75	0.80
	35	0.78	0.83	0.77	0.81
	40	0.79	0.84	0.78	0.83
	45	0.80	0.85	0.79	0.83
	50	0.81	0.86	0.80	0.84
	55	0.82	0.86	0.81	0.85
3	20	0.76	0.81	0.74	0.79

Occasions	Items	G		Φ	
		Sample 1	Sample 2	Sample 1	Sample 2
	25	0.79	0.83	0.77	0.81
	30	0.81	0.85	0.79	0.83
	35	0.82	0.86	0.81	0.85
	40	0.83	0.87	0.82	0.86
	45	0.84	0.88	0.83	0.87
	50	0.85	0.89	0.84	0.87
	55	0.86	0.89	0.85	0.88

In sample 2, the results are somewhat more favorable because the variance accounted for by real construct differences for people was higher (13%). This results in higher dependability coefficients. For the current design, the estimate for the G-Coefficient is .82 and the Phi-Coefficient is .80.

General Discussion

The present research investigated the score dependability of the Wisconsin Schizotypy Scales within the framework of G-Theory. The analyses pointed to similar conclusions for both samples and for all four scales. Here we will summarize the large number of findings and develop some conclusions for users of these scales.

First, the analyses consistently found that only small amounts of variance were accounted for by examinee differences in traits in both samples (9% in the Revised Social Anhedonia Scale, 6% in the Physical Anhedonia Scale, 10% and 14% in the Perceptual Aberration Scale, and 10% and 13% in the Magical Ideation Scale). Overall, then, the range of variance was 6% to 14%, which is, in a word, low. This suggests that real differences between examinees are measured poorly with these scales, as reflected by the score dependability coefficients. They are at or below .80 for the scales in the current design for the Physical Anhedonia Scale, the Perceptual Aberration Scale, and the Magical Ideation Scale; they are above .80 only for the Revised Social Anhedonia Scale.

Second, adding more than one time point to a research design would not be an efficient way to increase the dependability of the Wisconsin Schizotypy Scales. Occasions and their interactions don’t seem to have much influence on the variance of scores: they explained trivial amounts of variance. The small effects of occasions were reasonably stable from sample to sample, especially for the Revised Social Anhedonia Scale and the Physical Anhedonia Scale. They were less consistent over samples for the Perceptual Aberration Scale and the Magical Ideation Scale. Given the logistics of assessing a sample for more than one time point, few researchers using these scales would find multi-occasion assessment to be worth the modest payoff in dependability. This should be reassuring to test users, given that most research with these scales uses only one occasion of measurement.

Third, most of the variance was associated with items and with person-by-item interactions, and this fact provides direction for the future development of these scales. When extensive variance is due to items—the items perform differently from each other—researchers could typically improve a scale by adding items. For the Wisconsin Schizotypy Scales, however, this is probably not a worthwhile strategy. The D-studies estimated the influence of adding items on score dependability, and they showed that acceptable levels of dependability (.80 and greater) are often not reached or barely reached with large numbers of items. Adding more items to scales that are already long is inefficient.

Instead, these items need to be examined in more detail with an eye toward eventual scale revision. Item Response Theory (IRT) could offer valuable insight into the workings of the scale items. The large variance due to items indicates that some items behave differently than others. IRT analyses provide detail about each item's difficulty and discrimination levels and thus help identify unusual items. Identifying and removing items with poor discrimination levels, in particular, would be a good initial step in improving the scales' item quality.

Furthermore, the large variance due to the person-by-item interaction indicates that people with similar trait levels are responding differently to some items. In IRT terms, this represents differential item functioning (DIF). The large interactions are a clue that there could be widespread DIF in these scales. Methods for detecting DIF could help identify poorly-performing items for potential exclusion. By dropping high DIF items, researchers may be able to reduce the percent of variance due to items and their interactions and enhance the amount of variance due to examinees.

Regardless of the direction of future scale development—sanding down the scales based on IRT and DIF methods or starting from scratch—it seems clear that the Wisconsin Schizotypy Scales are not as dependable as they could be. The present findings are thus a good demonstration of how traditional reliability estimates, such as Cronbach's Alpha, can overstate the dependability of the scales because they account for only one source of error (i.e., error due to items). Considering several sources of error—and, critically, their interactions—reveals that aspects of the items carry a disproportionate influence relative to real trait differences between examinees.

It is important to consider some boundaries and limitations of the present findings. In particular, college students with a low base rate of schizotypy filled out the scales, and the sample sizes (160 and 102) are modest by the standards of large-scale assessment research. These features might have influenced the variance of scores in general because there were probably relatively few students with very high scores. Hence, the college student population was homogeneous, and it thus could contribute to the small percentages accounted for by examinee differences. A more meaningful measure of the scales' dependability might be provided by an analysis of a sample that included a larger representation of participants from a clinical population.

At the same time, we should note that the methods of G-theory are descriptive rather than inferential (Shavelson and Webb *1991*), and our samples resemble the kind of samples, both in composition and size, that many scale users will have. For example, probably most of the contemporary studies that use the Wisconsin Schizotypy Scales are concerned with subclinical

variation in scores, either as part of screening an initial population or examining subclinical variation in a non-clinical convenience sample. Furthermore, the congruence between our two samples suggests that the estimated variance components are stable.

The present results thus inform about how the scales will perform for these typical purposes, but readers should keep the nature of the sample in mind when using the present findings as a basis for planning future work. Scale users who intend to study much different participant groups, such as high-risk or clinical populations, should consider conducting and reporting generalizability analyses, which would provide valuable information on the dependability of these scales across different kinds of samples

References:

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Chapman, J. P., & Chapman, L. J. (1983). Reliability and the discrimination of normal and pathological groups. *Journal of Nervous and Mental Disease, 171*, 658–661.

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.

Chapman, L. J., Chapman, J. P., & Miller, E. N. (1982). Reliabilities and intercorrelations of eight measures of proneness to psychosis. *Journal of Consulting and Clinical Psychology, 50*, 187–195.

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, J. P., Chapman, L. J., & Kwapil, T. R. (1995). Scales for the measurement of schizotypy. In A. Raine, T. Lencz, & S. A. Mednick (Eds.), *Schizotypal personality* (pp. 79–106). New York: Cambridge University Press.

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.

Crick, G. E., & Brennan, R. L. (1983). *GENOVA* [Computer software]. Iowa City: The University of Iowa, Iowa Testing Programs.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks: Sage.

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

Eckblad, M. L., Chapman, L. J., Chapman, J. P., & Mishlove, M. (1982). *The revised social anhedonia* scale. Unpublished test, University of Wisconsin, Madison, WI.

Edell, W. S. (1995). The Wisconsin Psychosis-Proneness Scales. In G. A. Miller (Ed.), *The behavioral high-risk paradigm in psychopathology*. New York: Springer-Verlag.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: an introduction to generalizability theory. *Counseling Psychologist, 27*, 325–352.

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.

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.

McGlashan, T. H., Miller, T. J., Woods, S. W., Rosen, J. L., Hoffman, R. E., & Davidson, L. (2001). *Structured interview for prodromal syndromes*. New Haven: Yale School of Medicine.

Meehl, P. E. (1964). *Manual for use with checklist of schizotypic signs.* Unpublished manuscript.

Meehl, P. E. (1989). Schizotaxia revisited. *Archives of General Psychiatry, 46*, 935–944

Mishlove, M., & Chapman, L. J. (1985). Social anhedonia in the prediction of psychosis proneness. *Journal of Abnormal Psychology, 94*, 384–396.

Nugent, W. R. (2006). A psychometric study of the MPSI suicidal thoughts subscale. *Stress, Trauma and Crisis: An International Journal, 9*, 1–15.

Seeger Halvorsen, M., Hagtvet, K. A., & Monsen, J. T. (2006). The reliability of self-image change scores in psychotherapy research: an application of generalizability theory. *Psychotherapy: Theory, Research, Practice, Training, 43*, 308–321.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., et al. (2008). Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 68–85.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3*, 271–295.