Available online at www.sciencedirect.com
ScienceDirect
Thinking Skills and Creativity 3 (2008) 15–22
Thinking Skills and Creativity
http://elsevier.com/locate/tsc

Exploring the structure of the Watson–Glaser Critical Thinking Appraisal: One scale or many subscales?

Robert M. Bernard [a,*], Dai Zhang [a], Philip C. Abrami [a], Fiore Sicoly [a,b], Evgueni Borokhovski [a], Michael A. Surkes [a]

[a] *Centre for the Study of Learning and Performance, Concordia University, Montreal, Quebec, Canada*
[b] *Institute of University Partnerships & Advanced Studies, Georgian College, Barrie, Ontario, Canada*

Received 5 October 2006; received in revised form 22 September 2007; accepted 20 November 2007
Available online 21 February 2008

Abstract

Critical thinking (CT) has been of longstanding interest among scholars, educators, and others who are concerned with thinking skills. The Watson–Glaser Critical Thinking Appraisal (WGCTA) is the oldest and among the most widely used and studied CT measure. It was constructed around five subscales (or CT skills): inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments. This paper describes a two part analysis of the psychometric properties of the WGCTA, based on 13 sets of subscale inter-correlations and 60 sets of subscale means retrieved from published studies. We performed a meta-analysis on the inter-correlations of the10 combinations of subscales and found that all of the average correlations that resulted were significant, but that all but one was significantly heterogeneous. Subsequently, we conducted principal components analysis on 60 subscale means of two different versions of the WGCTA. Each produced a one-factor solution, accounting for 82.69% and 79.55% of the total variance, respectively. Together these two parts of this study suggest that the WGCTA should be viewed as a measure of general competency, and that the subscales should not be interpreted individually.

Keywords: Critical thinking; Measurements; Watson–Glaser Critical Thinking Appraisal; Psychometric structure

1. Introduction

Most educators would agree that learning to think is one of the most important goals of formal schooling. This includes not only thinking about important problems within disciplinary areas, such as history, science, and mathematics, but also thinking about the social, political, and ethical challenges of everyday life in a multi-faceted and increasingly complex world. Educators are not the only ones concerned about the urgency of prioritizing the teaching and learning of critical thinking (CT). In the United States, "a national survey of employers, policymakers, and educators found consensus that the dispositional as well as the skills dimension of critical thinking should be considered an essential outcome of a college education" (Tsui, 2002, pp. 740–741).

* Corresponding author at: CSLP, LB-581, Department of Education, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8. Tel.: +514 848 2424x2020.
E-mail address: bernard@education.concordia.ca (R.M. Bernard).

1871-1871/$ – see front matter
doi:10.1016/j.tsc.2007.11.001

Over the years, 13 standardized measures of CT have been developed and many are available for purchase. Scores on CT tests have been found to be predictive of or associated with success in a variety of settings (e.g., teaching critical thinking skills, Heraty & Morley, 2000; Wood, 1981; the success of educational programs, Sandor, Clark, Campbell, Rains, & Cascio. 1998; Dale, Ballotti, Handa, & Zych, 1997; teaching clinical skills, Miller, Sadler, & Mohl, 1993), and a variety of abilities (e.g., general academic success, McCammon, Golden, & Wuensch, 1988; Gadzella, Ginther, & Bryant, 1987; the ability to communicate with depressed patients, Gonzalez, 1996). Since prediction in these areas is largely dependent on the quality of the instruments used for measuring CT, it is important that we know something about their psychometric properties and especially what and how much they are capable of telling us.

Major critical thinking skill standardized tests, such as the Watson–Glaser CT Appraisal (WGCTA), the Cornell CT Test (CCTT), the California CT Skill Test (CCTST), and the Test of Critical Thinking-Form G, are usually devised around a set of subscales that collectively represent the "theory" or definition of critical thinking of the test as a whole. The WGCTA, for instance, contains items that are classified in terms of five CT subscales (or CT skills), deduction, recognizing assumptions, deduction, interpretation, and evaluating assumptions. Ideally, empirical validation of the instrument should at least roughly reveal the underlying theoretical constructs, and if it does not the question arises "What is the overall test actually measuring?" In addition, if there are distinguishable subscales, they should not be highly inter-correlated, suggesting that they are indeed measuring different aspects of the overall definition. According to Watson and Glaser (1980), the five subscales of the WGCTA were "each designed to tap a somewhat differing aspect" of CT skills (p. 1). We chose to examine the psychometric properties of the Watson–Glaser Critical Thinking Appraisal (WGCTA), as a case in point, because it is the oldest and perhaps the best known and certainly the most widely scrutinized of the CT standardized tests. The wealth of published data on it allowed us to investigate the psychometric properties of the WGCTA subscales in two different ways to determine if there are, indeed, separate and identifiable subscales, or if the test is best scored and interpreted as a general measure of CT.

In examining the literature, we found 13 published studies of the WGCTA that contained inter-correlation matrices for the five subscales. Within these, the inter-correlations varied widely across studies, from −0.07 to +0.74. Consistently low correlations, if their individual reliability estimates are reasonable high, suggest that the subscales are likely to be separate measures of the constructs. High correlations indicate the reverse, that the subscales are largely overlapping in what they are measuring.

1.1. Research questions

Our first research question relates to the 13 inter-correlation matrices from the literature: "What is the average correlation for each of the 10 pairs of inter-correlations among the subscales, and are these averages homogeneous (i.e., not variable enough to exceed what would expected from sampling error)." Procedures from meta-analysis were used to address these questions because samples containing correlations can be synthesized. The result of this analysis should shed some light on how highly inter-correlated the subscales are and whether the variability around the mean correlation exceeds the expectations of chance fluctuation.

As a follow-up to the first question, we also found 60 individual research studies that contained complete descriptive data (i.e., means) for each of the five subscales. Therefore, our second and interrelated question involved an assessment of the relationship of the five subscales to the entire test: When subjected to principal components analysis (PCA), are the five subscales somewhat separate from one another (i.e., multiple component structure), or do they cluster together as an indistinguishable unit (i.e., single component structure)?

In the test manual (1980), the authors of WGCTA do not encourage the use of the subscale scores for individual prediction or diagnosis, because the number of items in each subscale is small and therefore lacks sufficient reliability. But they claim that it is feasible to utilize the subscale scores to "analyze the critical thinking abilities of a class or larger group and to determine in light of such analysis the types of critical thinking training most needed by the group" (p. 9). In this study, we are investigating whether this claim is even warranted—Can the subscales in the WGCTA be discriminated from a psychometric perspective at the group level? If the average inter-correlations among the five subscales drawn for the 13 studies of the WGCTA is reasonably high (i.e., significantly higher than chance), then we expected that the PCA, conducted on independent data, would produce a one-component solution. The reverse would be expected to be true if the average inter-correlation among the 13 studies of the WGCTA were significantly less that expected by chance. For the first research question, we used statistical methods drawn form the literature of meta-analysis (Lipsey & Wilson, 2001; Hedges & Olkin, 1985). For the second research question, we used standard PCA.

The results of this study have implications beyond the use of the WGCTA as a research measure. Since the WGCTA is marketed and used as a predictive measure in a wide variety of educational and vocational contexts, and is sometimes used for the academic placement of individuals and for other diagnostic purposes, it would be useful to know if the individual subscales can be interpreted, or whether the general score is the only meaningful outcome.

2. Method

2.1. The Watson–Glaser Critical Thinking Appraisal

There are four standardized versions and one experimental edition of the Watson–Glaser measure. The first standardized version dates from 1951. The latest, the WGCTAs, was revised in 1994 and is a short form containing 40 items that can be administered in 45 min. The number of items varies across versions but the subscales and their descriptions have remained consistent over time. The WGCTA contains five subscales, labeled as: inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments. While not directly comparable to those of other tests, such as the California critical thinking skills test, they follow from the same theoretical underpinnings.

2.2. Retrieval of studies

Studies used as data resources in this research come mainly from two searches. We included relevant studies (primary studies), whose key terms in search strategies primarily included: "critical think*," "high order think*," "skills," "educat*," "teach*," "learn*," "control group," "compar*," "treatment," "test*," "evaluation," etc. We also sought reports of correlations of WGCTA subscales (correlational studies). The key terms used in searches were "subtest*," "subscore*," "factor*," "psychometric*," "correlat*," "validity." The studies came from resources that included electronic databases such as ERIC, PsycInfo, ProQuest Digital Dissertations, ABI/Inform Global on ProQuest, the Internet, major conference proceedings, reference lists, as well as the grey literature.

2.3. Outcomes of the searches and data extraction

Two raters were involved in the retrieval task, and they worked independently to assess the relevance of studies. We found 13 studies containing the full correlation matrices of the WGCTA subscales. The 60 studies containing the mean scores of WGCTA subscales came from both the primary study search and the correlational study search.

Data from pre-tests and post-tests were treated as two separate sets of subscale means, as were data from different groups in a study. Therefore, in all we collected 70 sets of subscale means (subsequently reduced to 60). We also gathered 14 correlation matrices of WGCTA subscales for different samples/groups (reduced to 13). Although we looked, we were unable to locate any inter-item correlation matrices in the published literature.

3. Results

3.1. Research question one: meta-analysis of subscale inter-correlations

A meta-analysis is a quantitative synthesis of effect sizes ($d+$) or correlations ($r+$) that is intended to provide a better estimate of population parameters than any single study can. The meta-analysis statistics reported here are from Hedges and Olkin (1985) and the data were analyzed using Comprehensive Meta-Analysis™ 2.0, a software package specialized for meta-analysis. This software generates the average correlation ($r+$), a z-test of $r+$, the 95% confidence interval of $r+$, and the test of homogeneity of $r+$ (Q_T). Q_T is a homogeneity statistic that is most commonly used in assessing a collection of effect sizes or correlation coefficients. When all findings share the same population correlation, Q_T has an approximate χ^2 distribution with $k-1$ degrees of freedom, where k is the number of effect sizes or correlations. If the obtained Q_T value is larger than the critical value, the findings are determined to be significantly heterogeneous, meaning that there is more variability in the effect sizes or correlations than chance fluctuation or sampling error would allow around a single population parameter. The $r+$ can still be interpreted but with much less certainty than if $r+$ were homogeneous.

Table 1
Average correlations ($r+$) for 10 combinations of subscales and their Q_T

Pairs of subscales	k	$r+$	z-test	95% confidence interval		Q_T
				Lower	Upper	
Inference/recognize assumptions	13	0.23	11.33*	0.19	0.27	31.82**
Inference/deduction	13	0.32	16.46*	0.29	0.36	35.87**
Inference/interpretation	13	0.31	15.62*	0.27	0.35	27.56**
Inference/evaluate assumptions	13	0.22	11.00*	0.18	0.26	21.82**
Recognize assumptions/deduction	13	0.27	13.73*	0.24	0.31	59.82**
Recognize assumptions/interpretation	13	0.25	12.31*	0.21	0.28	63.88**
Recognize assumptions/evaluate assumptions	13	0.17	8.41*	0.13	0.21	15.97
Deduction/interpretation	13	0.40	20.43*	0.36	0.43	44.93**
Deduction/evaluate assumptions	13	0.23	11.20*	0.19	0.26	52.54**
Interpretation/evaluate assumptions	13	0.31	15.56*	0.27	0.34	51.93**

*z (critical) = 1.96, (d.f. = 12), $p < 0.05$; **$\chi^2 = 21.03$, (d.f. = 12), $p < 0.05$.

We first calculated the homogeneity statistic Q_T for each individual correlation coefficient and identified one study as an outlier across all subscales (i.e., extreme Q_T). After the outlier was removed, we produced average correlations ($r+$) for each of the 10 combinations of subscales that contained 13 pairs of correlation coefficients from retrieved studies, and the Q_T for those average correlation coefficients.

The outcomes for the 13 correlations across the subscale pairs are shown in Table 1. All of the average correlations were significant and all but one (i.e., "recognize assumptions" correlated with "evaluate assumptions") was significantly heterogeneous. This pair also had the lowest $r+$. Overall the average correlations ranged from a low of 0.17 to a high of 0.40 and they are all positive. While not staggeringly high, in a small sample the majority could be judged to be of medium magnitude. The heterogeneous Q_T indicated that the observed average correlation coefficients are not consistent with the model of a single common population correlation parameter (Hedges & Olkin, 1985). In other words, the true population correlations could be higher or lower than the ones observed. Given the high variability around the correlation means, we judged the resolution of this research question to be inconclusive, although suggestive of the medium average correlations that would be expected if the subscales were intercorrelated.

3.2. Research question two: principal components analysis of subscale means

Since we were unable to find raw correlational data at the item level to conduct a factor analysis on aggregated data, we decided to take another approach and use the 70 sets of subscale means that we had found. We excluded eight sets because the data were incomplete and two others because data were from only a single version of the WGCTA. We decided to conduct separate PCAs on the two versions: WGCTA (Form Ym/Zm) and WGCTA (Form A/B) because: (a) different versions of the WGCTA contain different numbers of items in each subscale; (b) the available evidence about the equivalence of forms is from two study only; the correlation parameters between Form A and Ym is 0.78 in a study of 68 students, and 0.69 between Form B and Ym in a study of 122 students (Watson & Glaser, 1980).

Rust, Jones, and Kaiser (1962) provided a rationale for PCA on subscale means in lieu of items-level data. They claimed that according to the "Spearman–Brown" effect (Gulliksen, 1950, cited by Rust et al., 1962), the "inter-correlations among subtest scores should be higher than inter-correlations among individual item scores.

Table 2
Description of the two data sets

Test version	Test items	Data sets	N
WGCTA (Form Ym/Zm)	100	26	1019
WGCTA (Form A/B)	80	34	1867

Table 3
Description of the samples

Test version level	WGCTA (Form Ym/Zm)		WGCTA (Form A/B)	
	Frequency	Percent	Frequency	Percent
K-12	8	30.8	2	5.9
Undergraduate	13	50.0	27	79.4
Other adults	5	19.2	5	14.7
Total	26	100.0	34	100.0

This implies that PCA of subtest scores should yield stronger factors" (p. 254). Table 2 shows the number of test items, data sets and *N* of each version of the WGCTA. The number of cases involved in this study was large (see Table 2).

We found a limited amount of demographic data that was common across studies. However, we were able to describe the samples in terms of level of education. Table 3 shows a breakdown of the two test versions according to educational level/age.

Since the subscales in different versions contain different numbers of items, Table 4 shows the descriptive data for the WGCTA forms separately.

Two principal components analyses were performed on the subscale means weighted by sample size ($w = 1/\hat{\sigma}^2$). This procedure gives more weight to large samples as compared to small samples. The results of the analyses are shown in Table 5. We found a one-component solution in each of the two analyses, accounting for 82.69% (Forms Ym & Zm) and 79.55% (Forms A & B) of the total variance, respectively. The KMO for the first PCA was 0.855 and 0.839 for the second analysis. These statistics suggest that sampling was adequate. Deduction and interpretation were the highest loading subscales (reversed on the two versions) and assumption and evaluation were the lowest (also reversed). Overall, however, the two analyses are similar in many respects.

Table 4
Means, standard deviations and *N* of the two versions of the test

Test version subscales	WGCTA (Form Ym/Zm)			WGCTA (Form A/B)		
	Mean	S.D.	*N*	Mean	S.D.	*N*
Inference	11.08	1.28	1019	8.47	1.45	1867
Assumption	12.16	0.62	1019	10.88	0.94	1867
Deduction	18.69	1.58	1019	10.35	1.07	1867
Interpretation	17.84	1.73	1019	11.34	1.06	1867
Evaluation	10.12	0.71	1019	11.13	0.93	1867

Table 5
Results of subscale weighted means PCA (with component loadings)

Subscales	Test form	
	WGCTA (Form Ym/Zm)	WGCTA (Form A/B)
Eigenvalue (Factor 1)	4.135	3.977
Percentage of variance (Factor 1)	82.69	79.55
Inference (component loading)	0.911	0.898
Assumption (component loading)	0.825	0.892
Deduction (component loading)	0.968	0.942
Interpretation (component loading)	0.956	0.951
Evaluation (component loading)	0.880	0.765

4. Discussion

This project attempts to answer two interrelated research questions about the psychometric properties of the WGCTA, using independent data sets derived from the empirical literature. The first question ("What is the average correlation for each of the 10 pairs of inter-correlations among the subscales, and are these averages homogeneous (i.e., not variable enough to exceed what would expected from sampling error).") was addressed using a statistical approach to research synthesis called meta-analysis. We found that each of the 10 pairs of average correlations (i.e., each subscale paired with the others) was significant, but that all but one average correlation violated the assumption of homogeneity of correlation. This fact makes interpretation of the average correlations problematic.

The second research question was addressed using PCA (i.e., the first step in factor analysis). We found using 60 complete sets of subscale means, drawn from the literature, the one-component solution emerged for both forms of the WGCTA. This is further and more convincing evidence that over many samples of different learner groups, that no clear subscale structure is discernable. It is likely that these results are high tied to the large inconsistencies that have been observed in the reliability coefficients of the WGCTA (i.e., from 0.23 to 0.73 with an average of 0.47 across studies). It may be that the WGCTA is highly context specific, performing reasonably well with some learners in some settings and more poorly with other learners.

It is also possible that the WGCTA response format is flawed, thus leading to the wide-ranging, and somewhat low, reliabilities. Wagner and Harvey (2003) argue that the response format of the WGCTA may account for some of the reliability problems noted above. Specifically, the multiple-choice response format used in the WGCTA allows for high levels of successful guessing, such that, ". . . increased levels of successful guessing would be expected to increase the standard error of measurement and violate the assumption that measurement errors are random variables measured in mean deviation form; . . ." (p. 1). They further state that:

> The potential implications of this response format with respect to the likelihood of successful guessing are obvious: i.e., if guessing is random, and the correct alternative is evenly distributed, even examinees having very low true levels of critical thinking ability might be able to achieve a 50% success rate on approximately 80% of the WGCTA items (p. 1).

If successful guessing is a problem with interpretation of the general score on the WGCTA, as Watson and Harvey claim, it is further exacerbated at the subscale level where fewer items are involved. The result would be that subscale reliabilities are even more volatile and suspect than the general score.

The reliability problem may also result from the rather larger judgment component in the "inference" subscale, as Helmstadter (1985) has criticized. In many items of this subscale, the judgment component seems to "depend more on a personality response set related to how much evidence is required before one is convinced of an argument than on an ability to ascertain whether an inference is a valid one" (p. 1694).

While we are unable to speak definitively on matters of test construction and the reliability problem, the implications of our findings for research on CT and practical use of the WGCTA are important. From the research perspective, they suggest that we may need to abandon attempts to explore unique qualities or skills associated with critical thinking and concentrate instead on critical thinking as a collection of highly interrelated skills and abilities not easily divorced from one another, not operating separately, but existing conjointly and complementarily. On the other hand, perhaps we may need to develop new methods of evaluating critical thinking, which are sympathetic to views that posit the existence of separate sub-skills. The evidence we have examined to date suggests the former view is more likely.

The implications of these findings for practice are also important. If critical thinking is a general skill or one of highly related skills, approaches to instruction may need to focus more on developing the general qualities of thinking and less on specific sub-skills.

These results do not completely undermine the theoretical work of Facione (1990a,b) and others who posit the existence of finer-grained sub-skills, and they do not rule out the existence or importance of the CT dispositions hypothesized by Facione. What they likely do say is that discriminating among these sub-skills, from a psychometric perspective, is difficult to do.

While there is little doubt as to the desirability of teaching students to think critically, the concept of critical thinking, as we have examined it here, is a slippery one, both in terms of definition and measurement. It overlaps with elements of creativity (heuristics) and problem-solving (algorithmics) and is related to other areas of research on human mental activity, such as intelligence, meta-cognition, and self-regulation of learning. It is likely a mistake to think of CT as a

single entity, but at this time we think that it may be an even bigger mistake to attempt to parse it into a collection of identifiable and measurable components.

Acknowledgements

This study was supported by the *Fonds québécois de la recherche sur la société et la culture* and the Social Sciences and Humanities Research Council of Canada grants to Bernard and Abrami. The authors express appreciation to Anna Peretiatkovicz for her assistance and contributions.

References

Dale, P. M., Ballotti, D., Handa, S., & Zych, T. (1997). An approach to teaching problem solving in the classroom. *College Student Journal*, *1*, 76–79.

Facione, P. A. (1990a). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction, (executive summary). Millbrae, CA: California Academic Press. Retrieved July 22, 2002 from http://www.insightassessment.com/pdf_files/DEXadobe.PDF.

Facione, P. A. (1990b). The california critical thinking skills test—college level: Interpreting the CCST, group norms and sub-scores (technical report #4). Millbrae, CA: California Academic Press.

Gadzella, B. M., Ginther, D. W., & Bryant, W. (1987). Teaching and learning critical thinking skills. In *Paper presented at the 26th international congress of psychology* (ERIC Document Reproduction Service No. ED405313)

Gonzalez, E. W. (1996). Relationships of nurses' critical thinking ability and perceived patient self-dislosure to accuracy in assessment of depression. *Issues in Mental Health Nursing*, *17*, 111–112.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. NY: Academic Press.

Helmstadter, G. C. (1985). Review of the Watson–Glaser Critical Thinking Appraisal. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 1693–1694). Lincoln, NE: Buros Institute of Mental Measurements.

Heraty, N., & Morley, M. J. (2000). The application of the structure of intellect programme: A manufacturing facility experiment. *Journal of Managerial Psychology*, *15*, 691–711.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

McCammon, S., Golden, L., & Wuensch, K. L. (1988). Predicting course performance in freshman and sophomore physics courses: Women are more predictable than men. *Journal of Research in Science Training*, *25*, 501–510.

Miller, D. A., Sadler, J. Z., & Mohl, P. C. (1993). Critical thinking in preclinical examinations. *Academic Medicine*, *68*, 303–305.

Rust, V. I., Jones, R. S., & Kaiser, H. F. (1962). A factor-analytic study of critical thinking. *Journal of Educational Research*, *55*(6), 253–259.

Sandor, M. K., Clark, M., Campbell, D., Rains, A. P., & Cascio, R. (1998). Evaluating critical thinking skills in a scenario-based community health course. *Journal of Community Health Nursing*, *15*, 21–29.

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *The Journal of Higher Education*, *73*(3), 740–763.

Wagner, T. A., & Harvey, R. J. (2003). Developing a new critical thinking test using item response theory. In *Paper presented at the 2003 annual conference of the society for industrial and organizational psychology*.

Watson, G., & Glaser, E. M. (1980). *Watson–Glaser Critical Thinking Appraisal: Forms A and B*. San Antonio, TX: PsychCorp.

Wood, L. E. (1981). An "intelligent" program to teach logical thinking skills. *Behavior Research & Instrumentation*, *12*, 256–258.

Further reading

These references are studies in the datasets

Annis, L. F., & Annis, D. B. (1979). The impact of philosophy on students' critical thinking ability. *Contemporary Educational Psychology*, *4*(3), 219–226.

Chang, E. C. (1969). Norms and correlates of the Watson–Glaser Critical Thinking Appraisal and selected variables for Negro college students. (Doctoral dissertation, University of Oklahoma, 1969). *Dissertation Abstracts International*, *30*(5), 1860A.

Duckworth, J. B. (1968). The effect of instruction in general semantics on the critical thinking of tenth and eleventh grade students, (Ed.D. Dissertation). Detroit, MI: Wayne State University. (ERIC Document Reproduction Service No. ED040188).

Elliott, B., Oty, K., McArthur, J., & Clark, B. (2001). The effect of an interdisciplinary algebra/science course on students' problem solving skills, critical thinking skills and attitudes towards mathematics. *International Journal of Mathematical Education in Science and Technology*, *32*(6), 811–816.

Follman, J. C. (1969). A factor analytic study of three critical thinking tests, one English test, and one logical reasoning test. (Doctoral dissertation, Indiana University, 1969). *Dissertation Abstracts International*, *30*(3), 1015A.

Follman, J., Brown, L., & Burg, E. (1970). Factor analysis of critical thinking, logical reasoning, and English subtests. *Journal of Experimental Education*, *38*(4), 11–16.

Follman, J., Lowe, A. J., Johnson, R., & Bullock, J. (1972). Correlational and factor analysis of critical readings and critical thinking-fifth grade. In *Paper presented at the annual convention of the international reading association*.

Follman, J., & Miller, W. (1971). Statistical analysis of three critical thinking tests. *Educational and Psychological Measurements*, *31*, 519–520.

Follman, J., Miller, W., & Hernandez, D. (1969). Factor analysis of achievement, scholastic aptitude, and critical thinking subtests. *Journal of Experimental Education*, *38*(1), 48–53.

Frost, S. H. (1991). Fostering the critical thinking of college women through academic advising and faculty contact. *Journal of College Student Development*, *32*(4), 359–366.

Gadzella, B. M., & Masten, W. G. (1998). Critical thinking and learning processes for students in two major fields. *Journal of Instructional Psychology*, *25*(4), 256–261.

Gibson, J. W., Kibler, R. J., & Barker, L. L. (1968). Some relationships between selected creativity and critical thinking measures. *Psychological Reports*, *23*, 707–714.

Hicks, R. E., & Southey, G. N. (1990). The Watson–Glaser Critical Thinking Appraisal and the performance of business management students. *Psychological Test Bulletin*, *3*(2), 74–81.

Hunt, D., & Randhawa, B. S. (1973). Relationship between and among cognitive variables and achievement in computational science. *Educational and Psychological Measurement*, *33*(4), 921–928.

Inlow, F. H., & Chovan, W. (1993). Another search for the effects of teaching thinking and problem solving skills on college students' performance. *Journal of Instructional Psychology*, *20*(3), 215–223.

Keller, R. (1994). Effects of an instructional program on critical thinking and clinical decision-making skills of associate degree nursing students (nursing education). (Doctoral dissertation, University of South Florida, 1993). *Dissertation Abstracts International B*, *54*(9), 4601.

Landis, R. E., & Michael, W. B. (1981). The factorial validity of three measures of critical thinking within the context of Guilford's structure-of-intellect model for a sample of ninth grade students. *Educational and Psychological Measurement*, *41*(4), 1147–1166.

Loo, S. R., & Thorpe, K. (1999). A psychometric investigation of scores on the Watson–Glaser Critical Thinking Appraisal new form S. *Educational and Psychological Measurement*, *59*(6), 995–1003.

Lowe, A. J., Follman, J., Burley, W., & Follman, J. (1971). Psychometric analysis of critical reading and critical thinking test scores—twelfth grade. In F. P. Green (Ed.), *Reading: The right to participate—20th yearbook of the National Reading Conference* (pp. 142–147). Milwaukee, WI: The National Reading Conference, Inc.

McDonough, M. F. (1998). An assessment of critical thinking at the community college level. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 58*(7), 2561A. Columbia University Teachers College, New York. Retrieved May 25, 2005, from ProQuest Digital Dissertations database. (Publication No. AAT 9738997).

Mines, R. A., Hood, A., King, P., & Wood, P. (1990). Levels of intellectual development and associated critical thinking skills in college students. *Journal of College Student Development*, *31*(6), 538–547.

Norris, C., & Jackson, L. (1992). The effect of computer science instruction on critical thinking skills and mental alertness. *Journal of Research on Computing in Education*, *24*(3), 329–337.

Pascarella, E. T. (1989). The development of critical thinking: Does college make a difference? *Journal of College Student Development*, *30*(1), 19–26.

Rust, V. I. (1960). Factor analyses of three tests of critical thinking. *Journal of Experimental Education*, *29*, 177–182.

Stekel, F. D. (1969). Development of a more flexible physical science laboratory program for non-science majors with superior high school science backgrounds, (final report). Stevens Point, WI: Wisconsin State Universities Consortium of Research Development. (ERIC Document Reproduction Service No. ED053987).

Teixeira, K. (2002). An experimental study comparing critical thinking growth and learning styles in a traditional and workshop based introductory mathematics course. (Doctoral dissertation, New York University, 2002). *Dissertation Abstracts International*, *62*(10), 3327A.

Wallace, S. R., Thompson, T. E., & Wiegmann, B. A. (1994). The effect of preservice laserdisc presentation of question types and wait-time use on questioning and wait-time use in clinical experiences. (ERIC Document Reproduction Service No. ED383688).

Westbrook, B. W., & Sellers, J. R. (1967). Critical thinking, intelligence, and vocabulary. *Educational and Psychological Measurement*, *27*, 443–446.