The Development of the Culturally Responsive Teaching Assessment Instrument

Janet L. Applin
Western Kentucky University, janet.applin@wku.edu

Follow this and additional works at: http://digitalcommons.wku.edu/kte

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, Educational Methods Commons, Other Teacher Education and Professional Development Commons, and the Special Education and Teaching Commons

Recommended Citation
http://digitalcommons.wku.edu/kte/4

This Article is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Kentucky Teacher Educator by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.
The Development of the Culturally Responsive Teaching Assessment Instrument

Abstract
This article is concerned with developing an instrument to assess Culturally Responsive Teaching. A study was conducted to explore the process of developing a valid and reliable CRT Assessment Instrument. Teaching behaviors widely accepted to indicate culturally responsive teaching were operationalized and an observation instrument was developed to assess teachers' use of culturally responsive teaching in applied classroom settings. Teachers were observed using the instrument and it was evaluated for its reliability and validity. Results of the study indicated that the instrument was found to have acceptable inter-rater reliability for approximately half of the indicators. The results supported the content validity of the instrument but construct validity was lacking. Through this research, it was discovered that indicators for Culturally Responsive Teaching observation must be descriptive and physically based to obtain reliable and quantitative data.

Keywords
Culturally responsive teaching, CRT, assessment, diversity, teacher behaviors, evaluation, instruction, learning environment

Disciplines
Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Methods | Other Teacher Education and Professional Development | Special Education and Teaching

This article is available at TopSCHOLAR*: http://digitalcommons.wku.edu/kte/4
The Development of the Culturally Responsive Teaching Assessment Instrument
Thematic Topic: Diversity

Janet L. Applin, Assistant Professor
Western Kentucky University

Abstract

This article is concerned with developing an instrument to assess Culturally Responsive Teaching. A study was conducted to explore the process of developing a valid and reliable CRT Assessment Instrument. Teaching behaviors widely accepted to indicate culturally responsive teaching were operationalized and an observation instrument was developed to assess teachers' use of culturally responsive teaching in applied classroom settings. Teachers were observed using the instrument and it was evaluated for its reliability and validity. Results of the study indicated that the instrument was found to have acceptable inter-rater reliability for approximately half of the indicators. The results supported the content validity of the instrument but construct validity was lacking. Through this research, it was discovered that indicators for Culturally Responsive Teaching observation must be descriptive and physically based to obtain reliable and quantitative data.

In the past two decades, the number of racially, culturally, linguistically, and ethnically diverse students in American public schools has increased exponentially. In the year 2000, nearly one of every three Americans was African American, Hispanic or Latino, Asian American, or Native American (Council for Exceptional Children, IDEA 1997, 1999). Predictions for the future are that this trend will continue (Utley & Obiakor, 2001). Culturally Responsive Teaching, or CRT, has been proposed as one way to help educators meet the needs children from all backgrounds and increase student learning.

CRT is a combination of best teaching practices that take into account a child's cultural background, experiences, and strengths, and uses that information to: (a) arrange the classroom environment; (b) plan for and deliver instruction and, (c) to provide a safe and welcoming learning environment (Author, 2005). It is believed that CRT may address issues such as achievement gaps between different ethnicities and White students (Donahue, Voelk, Campbell, & Mazzeo, 1999; U.S. Dept. of Education, 2005; Utley & Obiakor, 2001); the need for more social justice in our educational systems (Bell, 1997); and the disproportionate representation of minority students in special education and gifted education programs (Artiles, Harry, Reschly, & Chinn, 2002; Artiles & Trent, 1994; Banks, 1999; MacMillan & Reschly, 1998; National Research Council, 2002;
Utley & Obiakor, 2001). Although the available literature provides many descriptions of what CRT should be (Author, 2005; Gay, 2000; 2002; Richards, Browne, & Forde, 2004), there are no available assessment instruments that operationally define the behaviors of CRT so that teacher educators may assess and intervene on these skills. In addition, there is scarce evidence to support a relationship between CRT and improved student achievement.

The purpose of this paper is to describe: (a) the procedures followed to develop a reliable and valid observation instrument to assess CRT; (b) the outcomes from piloting the instrument in applied classroom settings with practicing teachers; and (c) the limitations of the instrument and implications for teacher education programs and suggestions for the future use of this instrument. The ultimate goals in developing the CRT Assessment Instrument are to have a reliable and valid assessment instrument that allows teacher educators to evaluate their interventions to improve CRT and to be able to assess, in the future, the effects that CRT has on student achievement.

Method

Developing the CRT Assessment Instrument

The focus of developing this instrument was to achieve an assessment instrument that had content and construct validity and acceptable levels of inter-rater reliability. Chronbach and Meehl’s (1955) three steps to investigate the construct validity of a measure were used as a guideline in the development of the instrument. These three steps included: (a) articulating a set of theoretical concepts and their interrelations; (b) developing ways to measure the hypothetical constructs proposed by the theory; and (c) empirically testing the hypothesized relations among constructs and their observable manifestations.

Phase I: Articulating Theoretical Concepts of CRT

The theoretical concepts of CRT were articulated by thoroughly reviewing the literature and organizing it into categories of CRT. Gay’s (2000) five standards of CRT encompassed the overall consensus of the literature, albeit broadly, thus, five standards, plus one additional standard that was widely discussed in the literature, were used as a starting point. Indicators from Sobel, Taylor, and Anderson’s (2003) Diversity Responsive Teaching Observation Tool as well as the Council for Exceptional Children’s (CEC) Common Core of Knowledge and Skills for Multicultural Education which adhered to the six standards were then adapted and operationalized.

Sobel, Taylor, and Anderson’s (2003) Diversity Responsive Teaching Observation Tool was originally developed to assess the CRT of teachers in a Colorado school district. Skills from CEC’s Common Core of Knowledge and Skills in Multicultural Education (ME) and Special Education were also used to ensure that the standards and indicators had universality in their content and would be applicable to special education teachers as well as general education teachers. CEC skills and knowledge in all areas of special education teacher preparation are widely accepted as the standard of professional and ethical practice. The National Council for Accreditation of Teacher Education (NCATE), in their
Programs for the Preparation of Teachers, uses CEC guidelines to assess whether special education teacher education programs meet their rigorous standards for accreditation. The Interstate New Teacher Assessment and Support Consortium (INTASC) and their Special Education Sub-Committee are other organizations which rely heavily on CEC standards to guide their assessment of teacher education programs in special education. Several states use CEC standards as their benchmarks of teaching competency in special education as well. The standards that are required of teachers by CEC, NCATE, and INTASC are broad generalizations, difficult to measure, and needed to be operationalized for observation.

Phase 2: Developing Ways to Measure the Constructs

The six categories of CRT developed were: (a) Cultural Heritages; (b) School/Home Connections; (c) Instructional Strategies; (d) Interactions; (e) Curriculum/Materials; and (f) Personal Histories. A prototype assessment instrument was developed. The prototype CRT Assessment Instrument included the 6 categories with 3 to 5 operationalized indicators for each category for a total of 23 indicators. The instrument was then piloted in four classrooms and sent to experts in the field, including D. Sobel and S. Taylor (personal communication, May 2005), for feedback.

In the final CRT Assessment Instrument (See Appendix), skills were coded as either Yes, for observed or No, for not observed. This was in contrast to the prototype where a frequency count of each indicator was included. Teachers were considered to have a low rate of CRT if their raw score fell between 0 and 8; a medium rate of CRT if their raw score fell between 9 and 16; and a high rate of CRT if their score fell between 17 and 23. These ranges were chosen arbitrarily by breaking down the total raw score of observed indicators (23) into approximately even sections.

Phase 3: Empirically Testing the Relations among Constructs

The third guideline suggested by Chronbach and Meehl (1955) is to test the hypothesized relations among constructs and their observable manifestations. This was accomplished by using the instrument to measure the behaviors of CRT in applied classroom settings. As observational data were analyzed, it was determined which indicators from the CRT Assessment Instrument were most frequently observed or indicated and which indicators had acceptable levels of inter-rater reliability. Mean scores, median scores and ranges were calculated for the classrooms observed. Inter-rater reliability data was calculated using Kazdin’s (1982) point by point formula for 25% of the observations.

Setting and Participants

Setting

The CRT Assessment Instrument was utilized in 46 first, second, and third grade classrooms within four schools in a mid-sized city in the south. Permission to observe classrooms in the district for the purpose of this study was granted by the district superintendent and individual building principals. Teachers voluntarily agreed to be observed with the instrument and signed an informed consent form before
being observed. Demographic data from each school is presented in Table 1.

Classroom Teacher Participants

Forty-six certified elementary teachers were observed with the CRT Assessment Instrument. The majority of those observed was young, white, female, and had fewer than 10 years of teaching experience. Although the lack of diversity of those observed was disconcerting given the fact that the schools served diverse populations of students, it was not surprising or uncommon since the majority of the teaching force across the county is white, middle class, and female (Howard, 1999). Teacher demographics are summarized in Table 2 in Appendix B.

Procedures

Observational data were collected during one 30 to 45 minute observation of the participant teachers. Language Arts instructional time was targeted to reduce the variability in activities that might be observed. Language Arts activities were typically conducted in the mornings at each school setting. Specific observation times were not scheduled but teachers agreed to be observed during Language Arts times and indicated that they gave consent for the PI to observe during those times unannounced. During 25% of the observations, inter-rater reliability was checked with one of two trained observers in addition to the PI. Data from the observations were analyzed to evaluate the content and construct validity of the CRT Assessment Instrument using Haynes’ (2003) guidelines for content and construct validity as a benchmark. As is generally acceptable in social sciences research, those standards and indicators with inter-rater reliability of 80% or above were considered acceptable to conclude that inter-rater reliability was present.

Results

Raw Scores of Teachers

The highest raw score possible on the CRT Assessment Instrument was 23. This would have meant that all 23 indicators were observed during the assessment observation. The range of raw scores of the teachers observed was between 6 and 16. This means that 13% of the teachers observed scored in the low range in their use of CRT according to the CRT Assessment Instrument. The remaining 87% of those observed scored in the medium range of CRT with the mode being a raw score of 10. No teachers scored in the high range of CRT. The indicators from the categories of Instructional Strategies and Curriculum/Materials were those that were observed most frequently during the observations. The indicators from the categories of School/Home Connections and Personal History were those that were observed the least frequently during the observations (see Table 3).

Inter-Rater Reliability

The mean percentage of inter-rater reliability for each category ranged from 52.3% to 91.5%. The median percentage of inter-rater reliability for each category ranged from 50% to 83%. For the 23 indicators of CRT, the mean percentage of inter-rater agreement ranged from 17% to 100%. Twelve of the 23 indicators had a less than acceptable inter-rater reliability. Eleven of the indicators had adequate inter-rater reliability with percentages at 80% or above.
Six of those indicators had 100% inter-rater reliability. Table 3 presents inter-rater reliability data for each category and indicator as well as percentage of teachers observed to have exhibited each indicator.

Validity

Content Validity
The content validity of the CRT Assessment Instrument is believed to be strong because: (a) the indicators and standards of CRT were developed through a thorough review of the literature on CRT; (b) feedback from experts in the field was obtained and the instrument revised to reflect the feedback suggestions; (c) widely accepted skills of CRT from Gay (2002), Sobel, Taylor and Anderson (2003), and the CEC Common Core of Knowledge and Skills in Multicultural Education were adapted for the standards and indicators; and (d) the skills were operationalized for observation. According to Kerlinger (1986), a measure is said to have content validity if the substance or content of the measure is representative of the content or universality of the content of the property being measured. The content included in this instrument meets Kerlinger’s (1986) standard, as well as Haynes’ (2003) requirement that the elements be representative of and relevant to the targeted construct.

Construct Validity
Construct validity involves: (a) articulating a set of theoretical concepts and their interrelations; (b) developing ways to measure the constructs proposed by the theory; and (c) empirically testing the hypothesized relations among constructs and their observable manifestations (Chronbach & Meehl, 1955). Loevinger (1957) referred to these steps as substantive, structural, and external validity. For this instrument, the concepts and their interrelations were articulated through the use of universally accepted standards and indicators of CRT to ensure substantive validity. The development of how to measure the concepts of CRT was guided by an Applied Behavior Analysis framework through operationalizing the behaviors representative of Gay’s (2002) broad standards of CRT as well as others from the literature to ensure structural validity. However, empirically testing the hypothesized relationships between the standards and indicators lowered the construct validity of this instrument which affected the external validity of the measure.

Due to the fact that the inter-rater reliability coefficients for the majority of the indicators (12 out of 23) assessed were predominantly lower than the widely accepted standard of .80, the instrument, as a whole, cannot be said to have construct validity. Although there were some indicators that did have a high degree of inter-rater reliability (11 out of 23), to fully support the construct validity of an instrument, the psychometric properties of the instrument, such as inter-rater reliability, must be acceptable and the inter-rater reliability here cannot be ensured for a majority of the indicators.

The ultimate question when discussing the validity of the CRT Assessment Instrument must be, “does the instrument actually assess CRT?” Or, is it possible that the instrument assesses something else like generic best teaching practices?
Discussion

The primary purposes of this article were to describe the procedures undertaken to develop the CRT Assessment Instrument and to evaluate its inter-rater reliability and content and construct validity. The ultimate goal was to have a reliable and valid observation instrument to assess teachers' use of CRT methods so that the benefits of CRT on student outcomes could, eventually, be assessed. The results showed that the CRT Assessment Instrument had partial inter-rater reliability and support for content validity. The construct validity of the CRT Assessment Instrument was not supported through this evaluation.

Those indicators that had acceptable and better inter-rater reliability were more readily observed in one observation as they were indicators that were physically based, or descriptive variables, and could be seen in a variety of lessons and activities. This does not mean that the teachers exhibited these behaviors at higher rates, but that the observers agreed when they did or did not at a higher rate. Those indicators that had unacceptable inter-rater reliability did so, perhaps, because they were indicators that were more socially based, or inferential variables, than physically based or descriptive variables (Bakeman & Gottman, 1997; Gall, Gall, & Borg, 2003). Physically based or descriptive codes are not open to interpretation among observers as are inferential variables. In addition, the indicators may have been more evaluative which would require not only an inference on the part of the observer, but also an evaluative judgment. Based upon the data from the indicators, there was almost the same amount of acceptable inter-rater reliability as unacceptable. In light of this, the instrument cannot be said emphatically to have inter-rater reliability.

The second question of this study was to determine if an instrument developed to observe CRT would have content and construct validity. Criterion-Related and Concurrent Validity are important in behavioral assessment instruments, however, because there were no other instruments with which to compare, they were not the focus of the current study. Overall, this instrument is believed to have support for content validity but the construct validity was found to be lacking. Because the psychometric properties of this instrument did not fully support the validity, there is the possibility that the instrument may have been measuring something other than CRT even though the content of the instrument had universality. The relationship between CRT and best teaching practices overlap and some teacher behaviors included in this instrument are considered best teaching practices, period, without regard to CRT (Miller, 2002; Paulsen, 2005). Future versions of the CRT Assessment Instrument must establish better overall inter-rater reliability, and attempt to establish criterion-related validity, and concurrent validity to make the claim that the instrument is fully valid. Increasing the acceptable levels of inter-rater reliability by replacing those indicators that were more evaluative in nature with others that are more physically-based will also contribute to the construct validity.

Limitations of the Present Research

The lack of support for the inter-rater reliability and content and construct validity of the CRT Assessment Instrument may have been a direct result of the limitations of
the study. There was no similar observation instruments available with which to establish criterion related or concurrent validity which made it difficult to ensure the construct validity of this instrument. The number of teachers observed was only 46 which make generalizing any of the results unreasonable. Each teacher was only observed one time, which may have affected the raw scores. If the teachers had been observed more than once, indicators not observed the first visit may have been evident in subsequent observations, thereby increasing their raw scores. In addition, although teachers were all observed during Language Arts instruction, there was no control over the lessons being taught during the observations. While the limitations of the study may have affected the outcomes, they informed the development of future research with the CRT observation instrument.

Implications for Teacher Education & Professional Development and Future Research

Future research should focus on refining the instrument developed for this study. A reliable and valid instrument to assess CRT skills will allow deficits to be addressed through training. For this specific instrument, the indicators need to be examined and re-operationalized where needed for more reliable observations. Once the reliability is improved, the validity issues may be addressed.

If the CRT Assessment Instrument can be refined and revised to establish better reliability and validity, it can then be used as a measure to assess the best ways to train teachers to be more culturally responsive. Questions that need to be answered include: (a) what are the most effective strategies to increase the CRT of practicing and prospective teachers; (b) in what CRT skills are teachers most lacking; (c) in what CRT skills are teachers most skilled; and (d) how can this instrument be used to assess teacher growth and change? When those questions are answered, we can begin to provide evidence that CRT actually makes a difference in student achievement, social justice in education, and disproportionate representation issues.
References


