

DESIGNING AND VALIDATING A MEASURE OF TEACHER KNOWLEDGE OF
UNIVERSAL DESIGN FOR ASSESSMENT (UDA)

by

ELISA MEGAN JAMGOCHIAN

A DISSERTATION

Presented to the Department of Educational Methodology,
Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2010

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

Elisa Jamgochian

Title:

"Designing and Validating a Measure of Teacher Knowledge of Universal Design for Assessment (UDA)"

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Paul Yovanoff, Chairperson, Educational Methodology, Policy, and Leadership

Elizabeth Harn, Member, Special Education and Clinical Sciences

Leanne Ketterlin Geller, Member, Educational Methodology, Policy, and Leadership

Douglas Blandy, Outside Member, Arts and Administration

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

June 14, 2010

Original approval signatures are on file with the Graduate School and the University of Oregon Libraries.

© 2010 Elisa Megan Jamgochian

An Abstract of the Dissertation of
Elisa Megan Jamgochian for the degree of Doctor of Philosophy
in the Department of Educational Methodology, Policy, and Leadership
to be taken June 2010
Title: DESIGNING AND VALIDATING A MEASURE OF TEACHER KNOWLEDGE
OF UNIVERSAL DESIGN FOR ASSESSMENT (UDA)

Approved: _____
Paul Yovanoff, PhD

The primary purpose of this study was to design and validate a measure of teacher knowledge of Universal Design for Assessment (TK-UDA). Guided by a validity framework, a number of inferences, assumptions, and evidences supported this investigation. By addressing a series of research questions, evidence was garnered for the use of the measure to describe what teachers know about assessment accessibility issues through their application of seven UDA principles. The investigation used research designs and sampling procedures specific to each research question. The TK-UDA was designed to capture depth of knowledge, from background to declarative to applied, through a variety of item types. Internal, external, and teacher reviews provided evidence to support the content validity of the measure, and, based on the feedback from these reviews, the

measure was revised to improve content and clarity. The measure was then implemented online; a purposeful sample of experts and inservice and preservice teachers was invited to participate in the study. It was anticipated that these participants would represent a range of knowledge of UDA. Following measure implementation, analyses were conducted to evaluate whether performance on items accurately reflected a continuum of teacher knowledge. Evidence of discriminant/criterion-related validity was examined by evaluating group differences. Based on results from *t*-tests and MANOVAs, no significant differences between groups (based on level of expertise) were found. Item Response Theory (IRT) scaling of items along a continuum indicated that declarative knowledge items were generally less difficult than applied knowledge items. IRT scaling of person scores represented a rather narrow range of knowledge within the sample. Reliability estimates from the IRT scaling and test-retest indicated strong item reliability, relatively weak person reliability, and satisfactory test-retest reliability, respectively. To obtain evidence regarding the usefulness of the measure to determine professional development needs, a Kruskal-Wallis rank-order test was conducted to evaluate the differential difficulty of UDA elements within the applied knowledge section. This provided initial evidence for identifying professional development needs at the element level. These results provide information that will guide further instrument development and future research in this area.

CURRICULUM VITAE

NAME OF AUTHOR: Elisa Megan Jamgochian

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon • Eugene, OR

California State University, Fullerton • Fullerton, CA

University of California, Davis • Davis, CA

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2010, University of Oregon

Master of Science, Special Education - Mild/Moderate Disabilities, 2004,
California State University, Fullerton

Bachelor of Arts, Psychology, 2001, University of California, Davis

AREAS OF SPECIAL INTEREST:

- Teacher Preparation and Professional Development
- Assessment and Instructional Decision-Making
- Universal Design in Education
- Technology in Education

PROFESSIONAL EXPERIENCE:

Curriculum Development Assistant, Department of Educational Methodology,
Policy, and Leadership, University of Oregon, Fall 2009-Spring 2010

Supervised College Teaching, Department of Special Education and Clinical
Services, University of Oregon, Fall 2008-Spring 2010

Instructor, Department of Secondary Education, California State University,
Fullerton, Spring 2008

Research Assistant, Behavioral Research and Teaching, University of Oregon,
Fall 2007-Summer 2009

Practicum Supervisor, Department of Department of Special Education and
Clinical Services, University of Oregon, Summer 2007

Research Assistant, Center on Teaching and Learning, University of Oregon,
Spring 2007

Tutor, Support for Student Athletes, University of Oregon, Spring 2007

Teaching Assistant, Department of Teacher Education, University of Oregon, Fall
2006-Winter 2007

Resource Specialist, Fullerton School District, 2002-2006

GRANTS, AWARDS AND HONORS:

Graduate Teaching Fellowships (various), University of Oregon, 2006-2010

Doctoral Research Award, College of Education, University of Oregon, 2010

Travel Grant, Department of Educational Methodology, Policy, and Leadership,
University of Oregon, 2009 and 2010

PUBLICATIONS:

Ketterlin-Geller, L. R., & Jamgochian, E. M. (in press). Instructional Accommodations and Modifications that Support Learning. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Accessible tests of student achievement: Issues, innovations, and applications*. New York: Springer.

Ketterlin-Geller, L.R., Nelson-Walker, N.J., & Jamgochian, E. (2009). *Project DIVIDE: Instructional Module Development* (Tech. Rep. No. 09-01). Dallas, TX: Southern Methodist University, Education Policy and Leadership.

Jamgochian, E., & Ketterlin-Geller, L.R. (2009). *Project DIVIDE: Instructional Module External Review* (Tech. Rep. No. 09-02). Dallas, TX: Southern Methodist University, Education Policy and Leadership.

- Jamgochian, E. M., Harn, B. A., & Parisi, D. M. (2008). Characteristics of students who don't respond to research-based interventions. *CEC Today*. Available from: <http://www.cec.sped.org/AM/Template.cfm?Section=Search&Template=/CM/ContentDisplay.cfm&ContentID=10645>
- Jamgochian, E. M., Harn, B. A., & Parisi, D. M. (2008). Similarities and differences of students who don't respond to research-based interventions. *CEC Today*. Available from: <http://www.cec.sped.org/AM/Template.cfm?Section=Search&template=/CM/HTMLDisplay.cfm&ContentID=10675>
- Jamgochian, E. M. (2004). *Pre-referral Interventions and Internet Use by Teachers*. Unpublished master's thesis. California State University, Fullerton, CA.

TABLE OF CONTENTS

Chapter	Page
I. STATEMENT OF THE PROBLEM.....	1
II. LITERATURE SYNTHESIS	7
Origins of Universal Design.....	10
Universal Design in Education.....	11
Universal Design for Learning (UDL).....	12
Universal Design for Assessment (UDA)	15
UDA and Test Accommodations.....	19
Implications of UDA for Assessment Design	20
Practical Implications of UDA for Classroom Assessment	22
Measuring Teacher Knowledge of UDA	23
Validity Argument Framework for the TK-UDA Measure.....	25
UDA Exists.....	28
Teacher Knowledge of UDA Can Be Measured.....	29
Performance on Items Accurately Reflects a Continuum of Teacher Knowledge	30
Consequences of Score Use.....	33

Chapter	Page
III. METHODS.....	34
Measure Specifications	35
Is the Content of the Measure Representative of the Seven UDA Principles?	
Establishing Evidence of Content Validity.....	39
Design and Analysis.....	39
Participants	40
Measures	40
Procedures.....	40
Implementation of the TK-UDA Measure	41
Procedures.....	42
Participants	43
Measure	47
Does the Measure Yield Scores that Reflect a Continuum of	
Teacher Knowledge?	47
Analytic Procedures	47
Does the Measure Effectively Differentiate Levels of Expertise in Relation to (a)	
Teacher Knowledge of UDA (Overall) and (b) Types of Knowledge	
(Background, Declarative, Applied)? Establishing Criterion-Related	
Evidence.....	50
Analytic Procedures	50

Chapter	Page
Are UDA Element Domain Scores (Sub-Scores) from Applied Knowledge (Scenario) Items Useful for Identifying Professional Development Needs?.....	52
Analytic Procedures	52
IV. RESULTS.....	54
Evidence of Content Validity	54
Evidence that the Measure Yields Scores that Reflect a Continuum of Teacher Knowledge.....	56
Evidence of Reliability.....	68
Criterion-Related Evidence	71
Evidence Supporting the Use of the Measure for Identifying Professional Development Needs.....	73
V. DISCUSSION.....	79
Evidence of Content Validity	80
Evidence that the Measure Yields Scores that Reflect a Continuum of Teacher Knowledge.....	81
Evidence of Reliability.....	84
Criterion-Related Evidence	85
Evidence Supporting the Use of the Measure for Identifying Professional Development Needs.....	87

Chapter	Page
Consequences of Score Use and Considerations for Measure Revisions	88
Limitations.....	90
Limitations of Sample Size.....	90
Limitations of the Measure.....	91
Limitations of the Analyses.....	93
Directions for Future Research.....	94
APPENDICES.....	96
A. MEASURE BLUEPRINT	96
B. MEASURE OVERVIEW.....	98
C. RECRUITMENT EMAIL – INFORMED CONSENT.....	100
D. RECRUITMENT EMAIL – FOLLOW-UP.....	103
E. RECRUITMENT EMAIL - REMINDER.....	104
F. TK-UDA PART I.....	105
G. PART I CONTACT INFORMATION FORM.....	118
H. TK-UDA PART II.....	119
I. PART II CONTACT INFORMATION FORM.....	131
J. TK-UDA PART I INTERNAL REVIEW FORM.....	132
K. TK-UDA PART I INTERNAL REVIEW COMMENTS.....	149
L. TK-UDA PART II INTERNAL REVIEW FORM.....	154
M. TK-UDA PART II INTERNAL REVIEW COMMENTS.....	166
N. TK-UDA PART I EXTERNAL/TEACHER REVIEW FORM.....	167

Chapter	Page
O. TK-UDA PART II EXTERNAL/TEACHER REVIEW FORM.....	186
P. TK-UDA EXTERNAL/TEACHER REVIEW COMMENTS.....	205
REFERENCES.....	212

LIST OF FIGURES

Figure	Page
1. Validity argument framework for measuring and interpreting teacher knowledge of UDA.....	27
2. TK-UDA measure specification overview.....	36
3. Distribution of person scores and item difficulties.....	62

LIST OF TABLES

Table	Page
1. Search Terms for Literature, Research, and Examples	9
2. Universal Design for Learning (UDL): A Summary of Neural Networks, UDL Principles, and Learner Benefits.....	14
3. Participant Groups	43
4. Participant Demographics	44
5. Inservice Teacher Descriptives.....	45
6. Retest Participant Groups.....	46
7. External and Teacher Review: Aggregated Tallies	55
8. Descriptive Statistics for Background Knowledge Items	57
9. Descriptive Statistics for Each Section, Based on Percent Correct	58
10. Correlations between Performance on Background Knowledge Items and Declarative and Applied Knowledge Items	59
11. Descriptive Statistics from IRT Analyses (Based on Scale Scores).....	60
12. Items from Declarative and Applied Knowledge Sections Pertaining to Each UDA Element	63
13. Descriptive Statistics for Constructed Response Scenario Items	64
14. Scenario 1: Examples of Constructed Responses Coded by UDA Element.....	65
15. Scenario 2: Examples of Constructed Responses Coded by UDA Element.....	67
16. Descriptive Statistics for Test-Retest.....	70
17. Test-Retest Correlations.....	70

Table	Page
18. Descriptive Statistics Per Group for Each Section, Based on Percent Correct	72
19. Correlation Matrix for UDA Element Domain Scores	74
20. Kruskal-Wallis Test of UDA Element Scale Scores.....	75
21. Descriptive Statistics by Group for Each UDA Element.....	77

CHAPTER I

STATEMENT OF THE PROBLEM

Universal design is rooted in architecture and product design; at its core is the belief that products and environments can be designed “to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

Recent federal legislation has emphasized improving academic achievement for all students including students with disabilities, those who are economically disadvantaged, and English language learners (e.g., No Child Left Behind [NCLB], U. S. Department of Education [USDE], 2001; Title I of NCLB; Individuals with Disabilities Education Act [IDEA], 2004). Central to these regulations are efforts to close the achievement gap between educationally disadvantaged students and their peers, and increase access to and inclusion in general education curricula, as well as participation in educational accountability assessments. However, as noted by Meo (2008), “such laws do little to address the biggest impediment to improving student outcomes: the curriculum,

[including classroom-level assessments] which is often not flexible enough to enable teachers to meet the needs of diverse learners” (p. 22).

According to a compilation of data from state departments of education, the percentage of educationally disadvantaged students steadily increased between 2002 and 2006; the most recent data indicate 13.6% of students enrolled in public schools are students with disabilities, 40.9% are economically disadvantaged, and 8.5% are English language learners (Council of Chief State School Officers [CCSSO], 2008). Educationally disadvantaged students are often prevented from participating fully in instruction and learning and from demonstrating their knowledge and proficiency due to the multitude of skills and knowledge (including language fluency and cultural familiarity) required to approach and access information and assessments, some of which are irrelevant to the constructs being taught and measured (Dolan, Hall, Banjeree, Chun, & Strangman, 2005; Ketterlin-Geller, 2005; Coltrane, 2002). These construct-irrelevant, or “access” skills, prevent students not only from accessing course content, but also from demonstrating their proficiency in the domain tested and potentially undermine their performance (Dolan, et al, 2005). Consequently, test validity may be threatened, resulting in misguided interpretation and misinformed use of scores in decision-making (Messick, 1989). By removing access barriers, through appropriate accommodations and by designing instruction and assessments that incorporate principles of universal design, a wider range of students can effectively participate in learning and evaluation.

Traditional assessments (i.e., assessments that do not incorporate features of universal design), are limited to the extent that they exclude students “at the margins” (Dolan, Rose, Burling, Harms, & Way, 2007, p. 4), and assume similar expected

outcomes for a presumably homogenous group of students (Rose & Dolan, 2000). As is apparent from the diversity present in classrooms across the nation, these limitations provoke important considerations for the assessment of student achievement and its outcomes. For students ‘at the margins’, that is, “those students who are doing poorly in traditional classrooms and for whom assessment is often most important” (Dolan, et al., 2007, p. 4), traditional assessments are likely to be neither fair nor accurate (Rose & Meyer, 2002). In addition, the results of traditional assessments tend to be confounded by student characteristics (e.g., visual acuity, decoding ability, motivation) that are not intended elements of the construct being measured, thus interfering with accurate measurement and interpretation of student learning (Rose & Meyer, 2002; Ketterlin-Geller, 2008). For these reasons, a flexible approach to assessment that accurately measures and promotes logical interpretation of student performance is necessary to “enhance the meaningfulness of assessments for all students” (Dolan et al., 2007, p. 4). The concept and guiding principles of UDA hold the keys to improving student assessment.

As UDA and its applications continue to develop and evolve, and as classrooms become increasingly diverse, it is critically important for teachers to know the philosophy behind this concept, incorporate elements of UDA into their classroom assessments, and accurately interpret student performance and make instructional decisions based on universally designed tests. As knowledge brokers and assessors, teachers are responsible for implementing high-quality instructional and assessment practices. Although teachers themselves may be considered highly-qualified based on their content knowledge competence and possession of a teaching license (USDE, 2001), most teachers have had

limited training in ways in which to assess student learning beyond writing objectives and using traditional test and item formats (Ellwein & Graue [1996] as cited in Shepard, 2000; Stiggins, 1999). Knowledge and use of appropriate student assessment practices is essential to instructional and decision-making processes. As test developers, consumers, and instructional decision-makers, teachers need to look critically at existing measures of student achievement, their uses and implications, and inferences made from their results. As UDA elements are used to guide the development of assessments, from large-scale, high-stakes tests to those used to measure student performance at the classroom level, teachers and other educational stakeholders may reasonably anticipate better alignment across tests as a result of clearly defined and appropriately measured constructs and have greater confidence in the accessibility and accuracy of tests and subsequent student achievement outcomes. By ‘leveling the playing field’ at the classroom level, UDA supports more valid and accurate interpretations and comparisons of student performance.

Extending the concept and principles of UDA to classroom assessments will require addressing teacher knowledge in this area. To date, this appears to be uncharted territory. The first step in this endeavor, explicated in the following chapters, is to design and validate a measure of teacher knowledge of UDA. By addressing the following research questions, evidence is garnered for the use of the measure to describe what teachers know about assessment accessibility issues through their application of seven UDA principles (described in detail in the next chapter).

1. Is the content of the measure representative of the seven UDA principles?
2. Does the measure yield scores that reflect a continuum of teacher knowledge?
 - a. Is performance on background knowledge items correlated with performance on declarative and applied knowledge items?
 - b. Are declarative and applied knowledge scores correlated, forming a single UDA knowledge measurement dimension?
 - c. Are teachers' declarative and applied knowledge of UDA scores structured from high (declarative) to low (applied)?
3. Does the measure effectively differentiate levels of expertise, in relation to:
 - a. Teacher knowledge of UDA (overall)?
 - b. Types of knowledge (background, declarative, applied)?
4. Are UDA element domain scores (sub-scores) from applied knowledge (scenario) items useful for identifying professional development needs?
 - a. Are domain scores correlated, forming a single UDA skill measurement dimension?
 - b. Are domain scores differentially difficult?
 - c. Do domain scores differentiate experts from non-experts?

The results of this study primarily provide direction for measure revisions and further instrument development and some initial evidence that substantiates the need for teacher professional development in this area. In addition, this study sets the stage for future research that explores (a) the design and delivery of a professional development curriculum for UDA, (b) the use of the measure presented herein as a pre-/post-test to

evaluate the effectiveness of professional development programs in terms of increased teacher knowledge and application of UDA, and (c) specific applications of UDA to classroom assessments (including comparisons of student scores on UD and non-UD tests in various subject areas).

CHAPTER II

LITERATURE SYNTHESIS

The primary purpose of assessment is to evaluate student learning and progress, the results of which inform instructional practices. Assessments may also be used to evaluate the performance of teachers, schools, and districts; to make comparisons between schools, districts and states; and to evaluate the effects of changes in curricula or practice (Rose & Dolan, 2000). Given the many uses of assessment and its varied implications, from evaluation of student performance at the classroom level, to large-scale, high-stakes assessments that may determine a student's instructional placement, whether s/he is eligible for a diploma, or his/her ability to succeed in post-secondary education, the need for fair and accurate assessments is clear. Universal design for assessment (UDA) reduces sources of error that may interfere with the assessment of learning (Rose & Dolan, 2000), therefore yielding more accurate assessment results that lead to more appropriate and effective instructional decisions.

The philosophical roots of UDA are reflected in assessment standards and legal mandates. The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) state that:

...all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure. Just treatment also includes

such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth. In situations where individual or group test results are reported, just treatment also implies that such reporting be accurate and fully informative (p. 74).

Within NCLB (U.S. Department of Education, 2001) are provisions for testing at least 95% of the total student population and significant subgroups (averaged over three years; states determine subgroups). As noted by Secretary Paige (2004), “[p]articipation in assessments makes our schools more inclusive, responsive and fair in meeting the needs of struggling students, which is why accountability is at the heart of *No Child Left Behind*” (§3). The participation of a wider range of students with varying abilities, experiences, and linguistic backgrounds requires more flexible and accurate assessments (Dolan, et al., 2007). Johnstone (2003) notes that “[a]lthough much of the research conducted in UDA to date has been concerned with making assessments more accessible to students with disabilities, there is often a spillover effect for other students, that is, ...English language learners, struggling readers, and students from diverse socioeconomic backgrounds also benefit” (p. 169).

Although the provisions noted above allude to elements of universal design (described in detail below), the inclusion of UDA principles in large-scale assessments is still in its infancy. Application of UDA to classroom-level assessments is even less developed, but holds great potential and important implications for instructional practices. Embedding assessment into curriculum and instruction supports a formative cycle of ongoing feedback and decision-making that is critical to learning (Rose & Dolan, 2000). Rose and Dolan (2000) note that this type of evaluation is rarely done in schools,

and is often supplanted by summative evaluation which, the authors argue, often poses an “ultimate obstacle, hurdle, or failure detector” (¶ 28).

The purpose of this literature synthesis is to trace the theoretical and historical roots of universal design from its inception in architecture and product design to its applications in education, including universal design for learning (UDL) and universal design for assessment (UDA). A major focus of this synthesis is on UDA and its practical implications, including its impact on state and district policy and extension to classroom practices. This synthesis concludes with a discussion of the measure of teacher knowledge of universal design for assessment (TK-UDA) developed for this study and the validity evidences needed to support its uses and score interpretations. The literature, research, and examples cited throughout this paper were sought using the search terms listed in Table 1, and acquired primarily through the University of Oregon’s library (either electronically, or from journals housed in the university’s library, or through Summit – an inter-library loan system) or acquired through coursework. Additional articles and websites were accessed online through Google/Google Scholar.

Table 1

Search Terms for Literature, Research, and Examples

Concept	Alternative terms
Universal design	Universal design in education Universal design for learning (UDL) Universal design for assessment (UDA)
Educational assessment	Accountability/high-stakes assessments Classroom assessments/Assessment of student learning Accommodations/Test Accommodations

Table 1 (continued)

Search Terms for Literature, Research, and Examples

Concept	Alternative terms
Education policy	No Child Left Behind (NCLB, 2001) Individuals with Disabilities Education Act (IDEA, 2004) Title I (NCLB, 2001) U.S. school enrollment
Teacher knowledge	Espoused and enacted knowledge Teacher practice Teacher learning
Validity	Assessment/Test standards Validity framework Validity evidences – content, response processes, criterion, statistical analyses (reliability, model fit)

Origins of Universal Design

With its roots in architecture and product design, the intent of universal design is to benefit people of all ages and abilities by “making products, communications, and the built environment more usable by as many people as possible” (Center for Universal Design, 2008, ¶ 2). With the adoption of the Americans with Disabilities Act (ADA; 1990) and subsequent ADA Standards for Accessible Design (1991), public spaces began to change to improve physical accessibility (Center for Universal Design, 2008). Initially, changes in public spaces were designed as add-ons, the results of which were often “costly and unattractive” (Pisha & Coyne, 2001, p. 198). To address this issue, the term ‘universal design’ was coined by Ron Mace, an architect and wheelchair user, to promote

the idea that accessibility could be considered proactively within the design/development stages rather than as an afterthought (Pisha & Coyne, 2001).

The Center for Universal Design (CUD), founded in 1989 by Ron Mace, “is a national information, technical assistance, and research center that evaluates, develops, and promotes accessible and universal design in housing, commercial and public facilities, outdoor environments, and products”; the Center’s mission is to “improve environments and products through design innovation, research, education and design assistance” (CUD, 2008, ¶ 1). The CUD has established seven principles to guide the design of environments, products, and communications (www.design.ncsu.edu/cud). These include (a) equitable use, (b) flexibility in use, (c) simple and intuitive design, (d) perceptible information/effective communication, (e) tolerance for error, (f) low physical effort, and (g) appropriate size and space for approach and use. Although these principles primarily address design considerations for physical spaces, they have broad influence on other fields including healthcare, the arts, and education (Thompson & Thurlow, 2002).

Universal Design in Education

Adopting the universal design paradigm and adapting it to educational settings can promote effective inclusion of students, access to general education curricula, and assessment of student learning. According to Acrey, Johnstone, & Milligan (2005), “universal design is a philosophy that is applicable at the national, state, school, or classroom level” (p. 24). The President’s Commission on Excellence in Special Education (U.S. Department of Education, 2002) suggested collaboration between general and special education instructional systems to provide effective instruction in general education and specifically recommended incorporating universal design into

accountability tools. IDEA (2004) provides an additional impetus for universal design in education as it “mandates a fuller inclusion of individuals with disabilities in general education classrooms and activities” (Erlandson, 2002, p. 2). As noted by Rose and Meyer (2000), “Universal Design does not imply ‘one size fits all’ but rather acknowledges the need for alternatives to suit many different people’s needs” (§ 5). In educational contexts, the concept of universal design is applicable to both instruction and assessment. The following sections explicate Universal Design for Learning (UDL) and Universal Design for Assessment (UDA).

Universal design for learning (UDL). The ability for students to interact with curriculum and instruction is wholly dependent on their ability to access content in meaningful ways that promote learning (Orkwis & McLane, 1998). This is largely a condition of the design and flexibility of the curricular materials used in instruction. Orkwis and McLane (1998) define UDL as:

the design of instructional materials and activities that allows the learning goals to be achievable by individuals with wide differences in their abilities to see, hear, speak, move, read, write, understand English, attend, organize, engage, and remember. Universal design for learning is achieved by means of flexible curricular materials and activities that provide alternatives for students with disparities in abilities and backgrounds...Universal design does not mean that the instructional materials and activities accommodate students by lowering the standards. (p. 9)

In a universally designed curriculum, attention is paid to the goals of the learning experience (Rose & Meyer, 2000), materials and methods are appropriately challenging and flexible, and assessment is flexible, formative, and provides accurate information to help teachers make instructional decisions and maximize student learning (Hitchcock, Meyer, Rose, & Jackson, 2002). By recognizing, planning for, and supporting a continuum of student abilities, universally designed curricula include a variety of options for accessing, using, and engaging with information (Rose & Meyer, 2002); “UDL shifts the burden for reducing obstacles in the curriculum away from special educators and the students themselves and leads to the development of a flexible curriculum that can support all learners more effectively” (Hitchcock, et al., 2002, p. 9).

Researchers at the Center for Applied Special Technology (CAST) pioneered the concept of UDL and continue to study its applications and outcomes. Central to their UDL framework are three interconnected neural networks identified through cognitive neuroscience research – recognition, strategic, and affective – that address the “what”, “how”, and “why” of learning, respectively (CAST, 2008). Within the recognition network, objects and the overall context are discerned. The strategic network then promotes closer examination of objects and information to be gleaned. Finally, the affective network influences the length of time and amount of attention paid to the information. In order to support the roles of each brain network in learning, CAST researchers developed three UDL principles. These include: (a) multiple means of representation, (b) multiple means of action and expression, and (c) multiple means of engagement. Incorporating each of these into a universally-designed curriculum yields various learner benefits, including a variety of ways through which students can acquire

information, alternatives for learners to demonstrate their knowledge, and connections to learner interests (Table 2). Most importantly, inclusion of these principles promotes access to curriculum and instruction by reducing extraneous effort, often “expended in overcoming barriers and poorly designed pedagogies” (Hitchcock et al., p. 15).

Table 2

Universal Design for Learning (UDL): A Summary of Neural Networks, UDL Principles, and Learner Benefits

Neural network	UDL principle	Learner benefits
Recognition	Multiple, flexible means of representation	Gives learners various ways of acquiring information and knowledge
Strategic	Multiple, flexible means of action and expression	Provides alternatives for learners to demonstrate what they know
Affective	Multiple, flexible means of engagement	Draws on learners’ interests, presents appropriate challenges, and increases motivation

Adapted from CAST (2008); www.cast.org/research/udl

Rose & Meyer (2000) argue that “[a]lthough UDL would be theoretically possible using traditional materials, it is not practically feasible” (§ 23) due to logistical burdens of space, cost, and management. The authors contend that the use of digital multimedia technologies is ideally suited to UDL because of its versatility and flexibility. For example, a student reading a digital text has the ability to increase font size, hear text read aloud, click on a word to get its definition, and adjust the reading level (e.g., UDL Editions by CAST). It is important to recognize that multimedia tools are not inherently universally designed and can be as inflexible and inaccessible as print media; however, by considering the principles of universal design and embedding elements to support

learner interaction during software development, designers can avoid a number of barriers and promote access to content (Rose & Meyer, 2000; CAST, 2008).

Although Rose and Meyer (2000) doubt the practical feasibility of UDL without technology, it is important to note the value and impact of good pedagogy; that is, effective instructional design will certainly reflect UDL principles. Hitchcock et al. (2002) present a profile of a UDL classroom in which learning is fostered through multiple representations of content, models of skilled performance, scaffolded support, multiple and varied practice opportunities, and ongoing feedback, within a meaningful social environment that promotes collaboration over competition. These elements parallel many of those identified by Kame'enui and Simmons (1990; 1999) as elements of effective instructional design. The key to successful implementation of UDL lies in the acknowledgement that no single medium or method is accessible to all learners, and that the choice of content, media, and tools is intended to help students achieve learning goals through a balance of challenge and support (Hitchcock, et al., 2002).

Universal design for assessment (UDA). Critical to effective instruction is accurate assessment of student learning. Dolan and Hall (2001) state that “one of the most important and consequential elements of instruction is assessment. Whether assessment is embedded into teaching...or administered separately..., it must provide students with adequate and equitable means to express their knowledge and understanding if it is to provide accurate feedback on the performance of students” (p. 3). This sentiment is also endorsed by Menken (2000) and Coltrane (2002) who note the importance of alignment between classroom instruction, curricula, standards, and

assessment for accurate evaluation of student learning and effective instructional decision-making.

The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) state that the goal of standardized assessment is “to provide accurate and comparable measurement for everyone, and unfair advantage to no one. The degree of standardization is dictated by that goal, and by the intended use of the test” (p. 61). This presents a formidable, yet reasonable, challenge to assessment developers and teachers alike to ensure that student achievement and subsequent interpretations and decisions are based upon valid and reliable measures of students’ knowledge and skills (Johnstone, 2003). Universal Design for Assessment (UDA) extends the concept of universal design from the fields of architecture and product development to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002). UDA principles have recently been applied to large-scale and computer-based assessments to promote the participation of the widest range of students possible and valid interpretation of assessment results and student performance (Thompson et al., 2002).

The National Center on Educational Outcomes (Thompson et al., 2002), through a review of assessment, universal design, and instructional design literature, has identified the following seven elements of universally designed assessments:

1. Inclusive assessment population – Test development processes should consider the context of the populations to be assessed, including the range of abilities and skills within the population. Assessments should present appropriate opportunities for students to demonstrate their knowledge, and

need to be responsive to diversity, the inclusion of all students, and the demands of accountability.

2. Precisely defined constructs – Clearly defined constructs “are essential for making sound and valid educational decisions based on assessment results” (Johnstone, 2003). By clearly defining the construct to be measured and purpose of the assessment, construct-irrelevant barriers (i.e., cognitive, sensory, emotional, and physical obstacles) are reduced.
3. Accessible, non-biased items – Items are biased to the extent that they disadvantage a particular group of test-takers. Bias may result from the language of an item, such as words or phrases that are place or culture-specific, or may contain language that is insensitive to a particular gender or culture. Potentially biasing elements are defined by Popham and Lindheim (1980) as “anything in an item that could potentially advantage or disadvantage any subgroup of examinees within the populations to be tested” (cited in Thompson, et al., 2002, p. 10). In addition, measurement or item bias may be present if scores obtained by examinees who have the same ability, but are from different groups, yield different covariances among item responses (internal) or different correlations with non-test items (external). As a consequence, the measurement scale is varies across groups (scores are not comparable), or is differentially predictive, respectively.
4. Amenable to accommodations – Universally designed assessments may reduce, but not eliminate, the need for accommodations. Accommodations may include changes in test presentation, response format, time, and

environment to reduce the impact of a student's disability while maintaining the intended construct (Ketterlin-Geller & Johnstone, 2006). As a result, construct-irrelevant barriers are reduced, and access to test content is improved.

5. Simple, clear, and intuitive instructions and procedures – Directions and tasks should be understandable and consistent across sections of a test. An important consideration is whether or not students will be able to work independently through the assessment (Tindal & Fuchs, 1999).
6. Maximum readability and comprehensibility – Text and definitions should be simple and clear; content and important ideas should be presented in logical sequence. Conciseness and use of plain language do not alter the content, but instead improve comprehensibility and make content accessible to test takers.
7. Maximum legibility – Legibility refers to three main test features: text, illustrations, and response format. Text characteristics to be considered include contrast, type size, font, and spacing (between lines and letters). Illustrations, graphs and tables should support the content of the text and be clearly labeled; unrelated illustrations are unnecessary and often distracting (Johnstone, 2003). Black and white line drawings are the most clear. Response formats often require students to “bubble in” their answers. Generally, larger circles (“bubbles”) and allowing students to mark in their test booklet, rather than on a separate answer sheet, are recommended (Johnstone, 2003).

Together, these elements help to guide test developers in designing and improving assessments to meet minimum or baseline requirements for accessibility and to effectively measure the knowledge and skills of the widest range of students possible.

UDA and Test Accommodations

Universal design doesn't necessarily eliminate the need for accommodations, rather it sets the stage for ensuring accessibility to a broad range of students, some of whom may require additional changes to the assessment setting, presentation, response format, or timing to participate in assessment opportunities and demonstrate their knowledge and skills. Accommodations can be defined as "changes in instruction or assessment practices that reduce the impact of an individual's disability on his or her interaction with the material" (Ketterlin-Geller & Johnstone, 2006, p. 164).

Accommodations are intended to level the playing field by improving the accessibility of the test, not by altering the difficulty or construct (Tindal & Fuchs, 1999; Ketterlin-Geller & Johnstone, 2006). Typically, these are post hoc adaptations to the setting, presentation, response format, and/or timing of an assessment (Bremer, Clapper, Hitchcock, Hall & Kachgal, 2002). Thurlow et al. (2000) describe three requisite conditions for accommodations, including an established positive impact on student performance for students with disabilities, no impact for students without disabilities (i.e. the accommodation does not provide an advantage to students without the target disability), and maintenance of the measurement's psychometric properties.

Without accommodations, test validity may be compromised due to the interference of construct-irrelevant barriers with students' abilities to demonstrate their

knowledge and skills, effectively excluding them from participation in assessments (Dolan & Hall, 2001). Effective accommodations are those that reduce construct-irrelevant variance without changing the test construct (Ketterlin-Geller & Johnstone, 2006). To date, research on accommodations reveals varied effectiveness (e.g. Bremer et al., 2002; Johnstone, 2003). Accommodations are limited in a number of ways, including (a) variability in their assignment and administration across students, teachers (or test administrators) and settings, (b) restrictions in terms of what they can accomplish (Dolan & Hall, 2001), and (c) insensitivity to individual differences (Ketterlin-Geller, 2005). Technology can effectively support and standardize accommodations to reduce variability, promote independent access to test adaptations, and presents an effective tool for creating tests with embedded accommodations and elements of UDA (Dolan & Hall, 2001; Dolan et al., 2007; Johnstone, 2003).

Implications of UDA for Assessment Design

UDA has the potential to address the issues and limitations of accommodations and essentially reduce the need for test adaptations by “seek[ing] to amend the environment by creating individually tailored tests based on individual needs” (Ketterlin-Geller, 2005, p. 5). Although “accommodations can be an effective means for providing students with disabilities access to a test, they can only go so far in correcting assessments that test extraneous knowledge and abilities, such as reading abilities in a science test” (Dolan & Hall, 2001, p. 5). By embedding accommodations and support into assessments, rather than assigning them as add-ons to the test, students will be better

able to access test content and demonstrate their knowledge and understanding, teachers (and other stakeholders) will be able to more accurately compare student performance, and validity of educational decisions will improve (Ketterlin-Geller, 2005).

To develop a test that incorporates the elements of UDA, a number of considerations are necessary. At all stages of test development, Thompson, Johnstone, Anderson, and Miller (2005) recommend the following eight considerations: (a) incorporating elements of universal design in the early stages of test development, (b) including disability, technology, and language acquisition experts in item reviews, (c) providing professional development for item developers and reviewers on use of the considerations for universal design, (d) presenting the items being reviewed in the format in which they will appear on the test, (e) including standards being tested with the items being reviewed, (f) trying out items with students, (g) field testing items in accommodated formats, and (h) reviewing computer-based items on computers.

Additional considerations include content expert and stakeholder review of the assessment and the use of statistical procedures to determine item functioning. By including content experts in the review process, the test construct and content domain can be confirmed or refined, and the test can be reviewed for potential bias, readability and legibility, and suitability of materials and instructions (Hanna, 2005). Soliciting feedback from various stakeholders (e.g., students, parents, teachers, etc.) can reveal issues regarding the appropriateness and uses of the test in terms of the target population (Ketterlin-Geller, 2005). Statistical procedures, such as differential item functioning and item response theory, “ensure that the items accurately measure the intended construct

thereby generating meaningful data for decision making” (Ketterlin-Geller, 2005, p. 13).

These considerations, in conjunction with the UDA principles, provide a process by which appropriate and accessible tests can be developed.

Practical Implications of UDA for Classroom Assessment

Although UDA applications at the classroom level and its incorporation into curricula is in its nascency, the potential, applicability, and feasibility of UDA at the classroom level are illustrated in the following examples. Acrey et al. (2005) describe a three-step process for implementing UDA at the classroom level. In their study, teachers first became familiar with the philosophy of universal design through various readings, support from an outside consultant, and presentations from colleagues. Next, teachers developed study guides based on UDA elements, reviewed best practices in graphic design, and created a graphic design guide. Finally, teacher-created study guides were evaluated by colleagues and the research staff. Ultimately, this led to the formation of an on-site universal design team. Teachers reported better on-task student behavior and comprehension and increases in academic achievement as indicated by course grades.

Johnstone (2003) conducted a study to evaluate the differences in student performance on a traditionally designed mathematics test and one that incorporated elements of UDA. The traditional test was comprised of released state test items; for the universally designed version, these items were re-designed to remove construct-irrelevant information, bias, and time constraints, and improve clarity, accessibility, readability, and legibility. Tests were administered to students in a counter-balanced manner so that each student took both test types. The author also conducted interviews with 23 participants to gain insight into any perceived differences in each student’s own performance on the two

tests. Results reveal significant positive differences (i.e., higher achievement), for all students and subgroups included in the analysis, on the universally designed test, and an overall effect size of .39. Emergent themes from the student interviews reveal preferences for the UDA version of the test due to greater recognition of content, better readability, reduced test anxiety, and preference for responding directly on the test form. With the requisite use of large-scale assessments to measure student achievement, these results point to important considerations for the development and implementation of such tests, including the importance of training test designers to incorporate UDA principles and the potential of universally designed tests to better indicate student ability and knowledge.

The examples above illustrate the potential of UDA at the classroom level, an application that warrants further exploration, including, for example, further validation of the above findings and the effects of UDA in other content areas. Studies to date (e.g., Johnstone, 2003; Ketterlin-Geller, 2005; Dolan et al., 2005) have relied primarily on selected response items, thus research is also needed to evaluate other response formats. Importantly, as UDA principles continue to trickle down to the classroom, teacher training and professional development will need to address not only these changes in assessment practices, but also the roles of teachers as critical consumers and architects of tests and as instructional decision-makers.

Measuring Teacher Knowledge of UDA

As noted in the introduction, research that addresses teacher knowledge of universal design for assessment is lacking. Also notably deficient is teacher preparation for assessing student learning beyond traditional paper-and-pencil tests (e.g. multiple

choice, short answer, or essay; Stiggins, 1999). Given the importance of the instructional decision-making process, especially with respect to the role that assessment plays in higher-stakes decisions such as program placement and graduation, it is imperative that teachers have both knowledge of and the skills to implement quality assessment practices, including the ability to recognize poorly-designed elements of published tests in order to retrofit and/or assign accommodations to support student access and to design classroom assessments from the outset that incorporate minimum requirements for accessibility (such as the seven elements discussed above).

The measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA; Appendices F & H) was developed to evaluate practicing teachers' knowledge of test accessibility issues through application of UDA principles. The measure consists of four main sections: background, declarative, and applied knowledge, and demographic information. Items related to background knowledge are intended to assess a teacher's degree of familiarity with federal acts and regulations related to accessibility, experiences working with students of various backgrounds and abilities, provisions for student accommodations, and use of technology. Declarative knowledge of UDA is assessed via responses to a variety of true/false statements pertaining to the seven UDA principles identified by Thompson et al (2002). Applied knowledge is measured through test setting and example scenarios that present an assessment context and sample test question for which teachers are asked to evaluate examples and non-examples of each of seven UDA principles (discussed in the literature review above). Teachers are also asked to provide suggestions for revising two scenarios to improve their accessibility. Items pertaining to basic demographic information (e.g. grades, subjects, and years taught, educational

background) are also included. The purpose of the current study is to establish the validity of the use of this measure to describe teachers' knowledge of universally designed assessments.

Validity Argument Framework for the TK-UDA Measure

In test development and evaluation, the process of validation involves the logical explication of an interpretive argument that provides the rationale for the proposed uses and interpretations of a given measure (Kane, 1992; AERA, APA, NCME, 1999).

Validity, then, “refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests” (AERA, APA, NCME, 1999, p. 9; Messick, 1988).

Kane (1992) presents a framework for organizing and evaluating the inferences, assumptions, and evidences of the interpretive argument. Within this framework, the case for validation is made via an explicit, logical, and coherent argument that includes plausible assumptions (testable hypotheses), evidence that supports or disconfirms the assumptions, and reasonable conclusions. In particular, Kane presents a sequence of inferences, including observation, generalization, and extrapolation, that support the proposed interpretations and uses of a test score. Observation, or the “score result[ing] from an instance of the measurement procedure” (Kane, p. 529), is supported by procedural evidence, including, for example, test administration and scoring procedures. The score may be generalized to form inferences about performance on other, similar measures. Generalizations are supported by assumptions of invariance, that is, the conditions of measurement can vary without changes in outcomes. Evidence for generalization (i.e. consistency of scores) can be garnered from reliability or

generalizability studies. Within extrapolation, “conclusions are drawn about behavior that is different in important ways from that observed in the testing procedure” (Kane, p. 529). Such conclusions may be supported by evidence from qualitative analyses or criterion-related evidence. This validity argument framework is applied below to structure an argument for the use and interpretations of the measure of teacher knowledge of UDA developed for this study.

Figure 1 illustrates the proposed validity argument for the TK-UDA measure, from the score to its uses and interpretations at each level of inference. The discussion here is limited to the inferences and assumptions that pertain to the use of the measure for descriptive purposes (observation), with initial evidences provided for the usefulness of the measure to identify professional development needs. Additional propositions for evaluating the need for and effectiveness of professional development are included to illustrate potential additional uses of the measure and require additional validity evidences (e.g. expert review of the professional development modules, and pre-/post-test comparisons), but are beyond the scope of this study.

Three inferences are proposed to substantiate the use of the measure to describe teachers’ levels of knowledge of UDA. The first and second inferences suggest, respectively, that UDA exists and teacher knowledge of UDA can be measured. The third inference proposes that performance on items accurately reflects a continuum of teacher knowledge of UDA. Each inference is supported by assumptions that form testable hypotheses. The following sections are organized by these inferences and present proposed evidences to substantiate their respective assumptions.

Teacher Knowledge and Application of UDA			
Score	Inferences	Assumptions/Hypotheses	Evidence/Methods
<p>Observation</p> <p>Score Use: Describe teachers' levels of knowledge of UDA</p>	1. UDA Exists	1. 7 Elements of UDA are inclusive of all UDA principles	Literature review
	2. Teacher knowledge of UDA can be measured	2. Measure is representative of the 7 UDA Elements <ul style="list-style-type: none"> a. Background knowledge items are appropriate/ relevant b. Statements comprising 'declarative knowledge' items reflect a range of UDA principles c. Scenarios reflect a range of UDA principles and applications d. Constructed Response items appropriately extend application 	Content-related evidence Expert Review of: <ul style="list-style-type: none"> • Measure blueprint & overview • Content and clarity of measure Evidence based on response processes <ul style="list-style-type: none"> • Process study/structured online protocol
	3. Performance on items accurately reflects a continuum of teacher knowledge	3. Score is indicative of level of teacher knowledge of UDA	Evidence of discriminant validity/Criterion-related evidence
		4.1 Score is a reliable estimate of levels of teacher knowledge of UDA	Statistical Analyses (IRT; including reliability estimates; model fit)
		4.2 Scores for declarative and applied knowledge items differentiate types of knowledge	Test-retest
		5. Measure is sufficiently broad/captures low to high levels of knowledge	Statistical Analyses (IRT; items differentiate participants by overall ability level/knowledge)
<p>Generalization</p> <p>Score Use: Evaluate the need for professional development</p>	Level of knowledge of UDA is indicative of need for professional development in this area	6. Domain scores (UDA Element sub-scores) are useful for identifying professional development needs	IRT Scaling (Correlation, ranking) MANOVA (discriminant evidence)
<p>Extrapolation</p> <p>Score Use: Evaluate the effectiveness of professional development program</p>	Level of knowledge of UDA is indicative of effectiveness of professional development in this area		

Figure 1. Validity argument framework for measuring and interpreting teacher knowledge of UDA.

UDA exists. This first inference is supported by the assumption that the seven elements of UDA are inclusive of all UDA principles. This assumption is substantiated by theoretical evidence presented in the review of literature in the preceding sections, as well as the following summary. Together, these establish a basis for construct validation; that is, the “interpretation of a test’s properties or relations...decided by examining the entire body of evidence offered, together with what is asserted about the test in the context of this evidence” (Cronbach & Meehl, 1955, p. 284).

The seven elements of UDA identified by Thompson and colleagues (2002) resulted from the authors’ reviews of student assessment (including accommodations), universal design, and instructional design literature, and have since been used and cited in assessment and instruction research (e.g., Johnstone, 2003; Acrey, Johnstone, & Milligan, 2005) and guide large-scale assessment and curricula design. The authors argue that universally designed assessments “may reduce the need for accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves” (Thompson, Johnstone, Thurlow, p. 5). Given the potential of universal design principles to address the limitations of traditional assessments within the test development phase, the application of these seven elements may be considered *minimum* requirements for the design of tests that are accessible to the widest range of students possible. Designing tests in this way improves not only accessibility, but also increases accuracy in measuring student knowledge and skills, and, in turn, improves decision-making practices.

With this theoretical framework as a starting point, the assumptions and inferences that follow become both grounded in and guided by research. The construct representativeness of the measure can be supported by expert reviews of the measure blueprint and content, described below.

Teacher knowledge of UDA can be measured. The second inference is at the heart of this study, and is supported by an overarching assumption that the measure presented herein accurately reflects teacher knowledge and application of the seven elements of UDA. To substantiate this assumption, expert reviews of the measure blueprint and content (i.e., items and response formats) and a study of teachers' response processes are essential.

Expert reviews provide evidence for the appropriateness and representativeness of the test content, that is, of “the relationship between parts of the test and the construct” (AERA, APA, NCME, 1999, p. 11), as well as the clarity of the measure. The expert review is guided by four sub-propositions: (a) background knowledge items are relevant and appropriate, (b) statements comprising ‘declarative knowledge’ items reflect a range of UDA principles, (c) scenarios reflect a range of UDA principles and applications, and (d) constructed response items appropriately extend application of UDA elements. To verify each of these propositions, two groups of experts, internal and external, reviewed the measure overview and blueprint, which provided a visual of the measure content and representativeness of each of the seven elements of UDA, and the measure itself to evaluate content and clarity with regard to the sub-propositions noted above. This process is especially important as “[t]he appropriateness of a given content domain is related to

the specific inferences to be made from test scores” (AERA, APA, NCME, 1999, p. 12). Each of the expert reviews contribute to measure revisions, ideally yielding a measure that more accurately represents the construct.

To further substantiate the ultimate use of the measure to describe teachers’ knowledge of universally-designed assessments, evidence can be garnered from the response processes engaged in by teachers as they complete the measure. Using a structured online protocol (similar to a verbal protocol), teachers were asked to describe their approaches to and processes for responding to items, including, for example, what misinterpretations might arise from the wording of each item. Questioning teachers as they complete items provides insight into their interpretations of the measure’s content. This information, aggregated with the external expert reviews, was used to improve the clarity of the items, and contributed to a more accurate measure of teacher knowledge.

Performance on items accurately reflects a continuum of teacher knowledge.

In order to describe teachers’ levels of knowledge (the proposed score use for this study), performance on the measure should reflect a range of teacher knowledge. The major sections of the measure are intended to reflect a continuum of teacher knowledge from background (or emerging) to declarative to applied. This inference is supported by three main assumptions that can be evaluated using statistical analyses.

After the review procedures outlined above were completed, the measure was implemented. The process of implementing the measure (described in detail in the Methods section) involved a purposive sample of experts, teachers, and non-experts (pre-service teachers) completing an online version of the measure that captured responses electronically. Results were analyzed to evaluate the assumptions that (a) a score on the

measure is a reliable estimate of levels of teacher knowledge of UDA, (b) scores differentiate types of knowledge, and (c) the measure captures low to high levels of knowledge. Using a bi-factor model (Gibbons and Hedeker, 1992) to evaluate item-level information dependent on a common stimulus (particularly relevant to the test setting and item scenarios included in this measure) is most appropriate for representing factorial structures for measures that have a general factor (teacher knowledge of UDA), and specific factors (evaluating accessibility/UDA within a given scenario). In addition, items from the declarative knowledge section was scaled and evaluated for differential difficulty.

To substantiate the assumption that a score resulting from the measure is indicative of level of teacher knowledge of UDA, it is necessary to first establish criterion-related evidence. Since measures of similar constructs against which relevant criterion might be evaluated do not yet exist, differences in group performance on the measure, or discriminant validity, can be used to test the hypothesis that scores are indicative of different levels of proficiency. According to the *Standards*, “[c]ategorical variables including group membership variables, become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test interpretation is to be supported” (AERA, APA, NCME, 1999, p. 13). The sample of participants who took the measure included UDA experts and non-experts. To evaluate the significance of group differences, a *t*-test was conducted. It was anticipated that differences in levels of proficiency (high versus low) would exist between the expert and non-expert groups. In addition, a multivariate analysis of variance (MANOVA) was used to further evaluate group differences for the declarative and

applied knowledge types. If this assumption is not supported, measure revisions and subsequent evaluation of group differences are necessary.

Reliability estimates and model fit statistics indicate the degree to which (a) the measure can effectively discriminate levels of ability and (b) the fit of the data to the proposed model. In addition, test-retest reliability can be estimated by administering the measure to a subgroup of the original sample of participants at a later time. Correlation between these sets of responses is indicative of the reliability of the measure.

Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores...[and] the consistency of classifications of individuals derived from the scores. To the extent that scores reflect random errors of measurement, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision-making is limited (AERA, APA, NCME, 1999, p. 31).

Together, these analyses provided evidence for the use of the score for describing teacher knowledge of UDA.

Additionally, element domain scores provided initial evidence for the usefulness of the measure for targeting professional development at the level of UDA element. Scores were sampled from the applied knowledge section by element, then ranked and evaluated to determine the differential difficulty of the UDA elements. In addition, participant scores for each element were compared by group (experts, inservice, and preservice teachers). These results provide information for the potential utility of developing professional development modules targeted at the domain (UDA element) level.

Consequences of score use. The proposed use of the measure for this study was to describe teachers' knowledge of universal design for assessment and provide initial evidence for its usefulness in identifying professional development needs at the UDA element level. The validity argument outlined above describes a chain of inferences, assumptions, and evidences intended to support the use of the measure for descriptive purposes. However, the reliability of this framework relies heavily upon the information garnered at each stage of the evidentiary process; that is, inferences and assumptions may be upheld or refuted based on the results of the methods noted above. As such, the validity argument is dynamic and subject to change (Kane, 1992). Any additional uses or interpretations of the test score in decision-making, including those noted in the generalization and extrapolation stages (Figure 1), require further validity evidences. The methods section that follows outlines the processes and procedures for gathering and analyzing evidences in relation to specific research questions that align with the assumptions presented in this framework.

CHAPTER III

METHODS

The purpose of this study was to validate the use of the measure of teacher knowledge of universal design for assessment (TK-UDA), by providing evidence for the inferences and assumptions described previously, to describe what teachers know about assessment accessibility issues through their application of seven UDA principles. The methods combine descriptive, scaling, and statistical procedures to address the following research questions.

1. Is the content of the measure representative of the seven UDA principles?
2. Does the measure yield scores that reflect a continuum of teacher knowledge?
 - a. Is performance on background knowledge items correlated with performance on declarative and applied knowledge items?
 - b. Are declarative and applied knowledge scores correlated, forming a single UDA knowledge measurement dimension?
 - c. Are teachers' declarative and applied knowledge of UDA scores structured from high (declarative) to low (applied)?
3. Does the measure effectively differentiate levels of expertise, in relation to:
 - a. Teacher knowledge of UDA (overall)?

- b. Types of knowledge (background, declarative, applied)?
4. Are UDA element domain scores (sub-scores) from applied knowledge (scenario) items useful for identifying professional development needs?
 - a. Are domain scores correlated, forming a single UDA skill measurement dimension?
 - b. Are domain scores differentially difficult?
 - c. Do domain scores differentiate experts from non-experts?

The investigation used research designs and sampling procedures specific to each research question. The measure specifications that follow detail the content, response formats, and scoring procedures of the TK-UDA measure. Then, through a series of expert and teacher reviews and analyses of participants' scores, evidence was garnered for content- and criterion-related validities, the measure's usefulness for differentiating levels of teacher knowledge along a continuum and for identifying professional development needs.

Measure Specifications

The purpose of the measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA) is to evaluate teachers' knowledge of test accessibility issues through application of the seven elements of UDA described in the previous chapter. In general, the measure's content was based upon and derived, in part, from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999) to reflect a continuum of knowledge from background to applied. The overall structure of the measure is illustrated in Figure 2 (see also Appendix B). The

measure, TK-UDA, is comprised of four main sections, each of which is described in detail below.

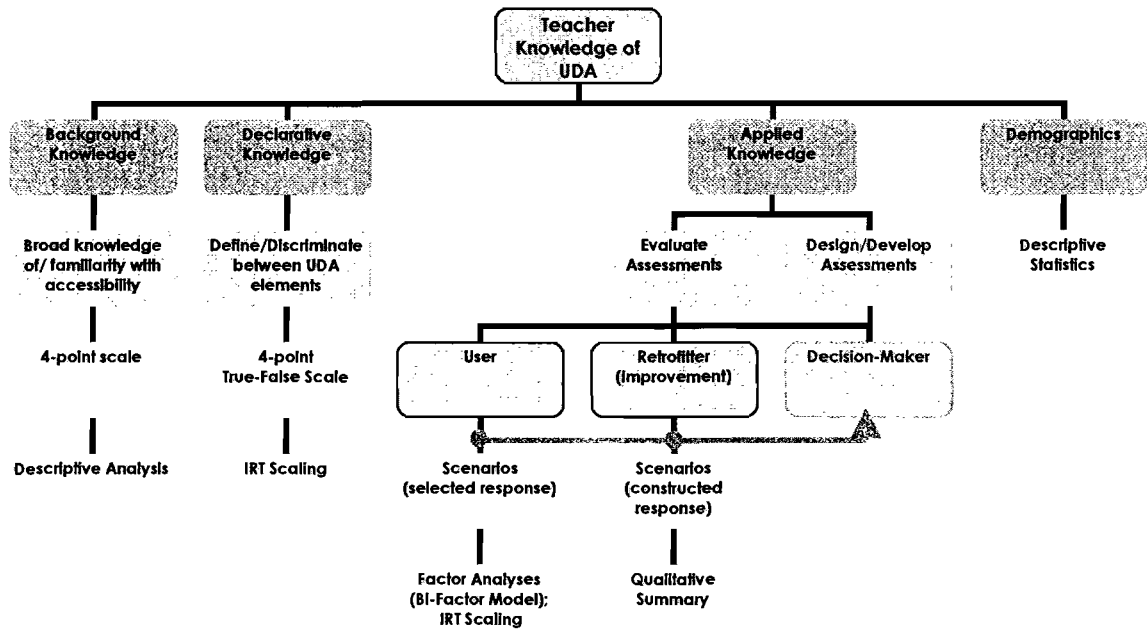


Figure 2. TK-UDA measure specification overview.

The first section, background knowledge, was comprised of 36 items intended to measure each participant's familiarity with federal acts and regulations related to accessibility, his/her experiences working with students of various abilities and backgrounds, provisions for allowing student accommodations, and uses of technology. Participants rated each item on a 4-point scale (e.g., 'not at all', 'somewhat', 'mostly', 'very'). Items were scored from 1 to 4 (low to high). Two items had follow-up questions for which teachers were asked to 'check all that apply' and/or fill in a blank to describe 'other'. For each of these, items were tallied if checked.

In the next section, declarative knowledge, participants were presented with 20 statements that reflect declarative (or factual) knowledge of the elements of UDA. The content for these statements was based upon descriptions of each of the seven UDA elements found in current research. For each of these statements, participants responded on a 4-point true-false scale (i.e., ‘very true’, ‘somewhat true’, ‘somewhat false’, ‘very false’). The scale was created in this manner in an attempt to prevent participants from skipping items they may have been uncomfortable stating as absolutely true or false. Items were scored correct (1) or incorrect (0).

Applied knowledge was conceptualized here as two skill areas: (a) evaluation of existing assessments and (b) design and development of new assessments. Teachers’ roles in this context were defined as user, retrofitter, and decision-maker. However, the design and development of accessible classroom assessments and the role of teacher as decision-maker were beyond the scope of this study, and therefore, were not addressed in the measure. To measure applied knowledge (i.e., teachers’ skill in evaluating assessments as users and retrofitters), participants were first presented with six scenarios that provided a description of a test setting and a sample student test item. Such context-dependent items are often considered “more realistic and perhaps even better for measuring higher-level skills” than single, independent items (DeMars, 2006, p. 145). For each of the given scenarios, participants evaluated the context (test setting and sample item) for accessibility using as their criteria the seven UDA elements. All student test items included in the scenarios were actual test items obtained from tests or student study materials available online. Participants responded ‘yes’ if a positive example of the element was presented in the scenario (it is accessible), ‘no’ if a negative example was

presented (it is not accessible), or 'N/A' if the element was not described in the scenario (it is not applicable). For each element within a scenario, responses were scored correct (1) or incorrect (0). Next, participants were presented with two additional scenarios for which they were asked to describe how they would revise the scenario to improve its accessibility with regard to the test setting, directions, and sample item. The two constructed response items represented a range of UDA elements and extended the application of UDA to address the role of the teacher as 'retrofitter'. Given the number of selected response items, the projected amount of time participants need to complete the measure (approximately 40 minutes), and time required for scoring these items, two constructed response items were deemed sufficient to extend teacher application. Responses to these items were tallied by UDA element (i.e., participant comments that identified or alluded to an element that needed improvement were counted; one tally for each element identified per scenario). These items were also evaluated qualitatively to illustrate common themes.

Lastly, items pertaining to basic demographic information (e.g. grades, subjects, and years taught, educational background) were analyzed descriptively.

The measure was created and delivered via a web-based interface (www.questionpro.com). This supported timely dissemination of the measure and data collection that was less cumbersome and more efficient than with a paper-and-pencil version of the measure. Studies comparing paper-and-pencil to computer progressive tests generally yield comparable scores (e.g., ODE, 2007). In addition, administration of the measure was standard across browsers. Data from full completion, partial, and multiple attempts were captured, as well as information regarding date, time, and

duration, and were downloaded as Excel files. The most complete data file for each participant was included in the analyses. Although the measure was untimed, participants were unable to save and return to their attempt, as the interface did not support this option. It was expected that participants would need approximately 40 minutes to complete the measure. Because of the limitations of the survey system and the anticipated time required for participants to complete the measure, the TK-UDA was delivered in two parts. Part I included sections for background and declarative knowledge, and demographics; part II included the applied knowledge test scenarios.

Is the Content of the Measure Representative of the Seven UDA Principles?

Establishing Evidence of Content Validity

To establish evidence of content validity, reviews of the measure blueprint (i.e., representation of the seven UDA elements across the espoused and enacted knowledge sections; Appendix A), measure specification overview (Figure 2; Appendix B), and the measure itself, provided evidence based on test content, as well as contributed to improved clarity of the measure (AERA, APA, & NCME, 2009). Four sub-propositions of the validity argument framework (Figure 1) guided this review: (a) background knowledge items are relevant and appropriate, (b) statements comprising ‘declarative knowledge’ items reflect a range of UDA principles, and (c) scenarios reflect a range of UDA principles and applications, and (d) constructed response items appropriately extend application of UDA elements.

Design and analysis. A series of internal, external, and teacher reviews provided evidence for content validation. At each phase, participant comments were summarized

and qualitatively evaluated for emerging and converging themes. Following each review, the measure was revised to improve content and clarity.

Participants. Three groups of purposefully selected participants reviewed the measure. First, an internal review was completed by three researchers at the University of Oregon with an interest in educational assessment and/or UDA. Next, three experts, namely, researchers with interests and research experiences in assessment and/or Universal Design in education, provided an external review. Last, a group of three teachers were asked to review the measure to further ensure clarity and consistent interpretation of items.

Measures. All participants completed an online review that included the measure (described above), as well as additional fields for reviewer comments. Following the internal review, additional items were included to obtain more specific feedback about each item on the TK-UDA from the external and teacher reviewers. (See Appendices J, L, N, and O for copies of review forms). Specifically, external and teacher reviewers were asked to rate and comment on the clarity of directions and items, as well as the appropriateness of the response scales used. This provided additional evidence for the appropriateness and clarity of the items, as well as evidence for response processes by highlighting potential misinterpretations. Internal and external reviewers were also provided with an electronic copy of the measure specification overview and blueprint (Appendices A and B).

Procedures. Participants were emailed a request to take part in the review, which included a brief description of the project, measure, and purpose of the review. Those interested were emailed a link to the online measure. Internal and external reviewers also

received the measure specification overview and blueprint for their review. Reviewers could elect to provide comments for the measure specification overview and measure blueprint either by ‘tracking changes’ electronically in the Word document, or by commenting directly on the paper copy. Reviews could be returned via email, mail, or in person. The measure, measure specification overview, and blueprint were first reviewed internally; suggested revisions pertaining to improving the content and clarity of the measure were made. The same process was used for the external/expert review. For the teacher review, participants completed only the online measure review. Participants were asked to complete their reviews within a two-week period; reminder emails were sent at the end of the first week to encourage those who had not already completed their review to do so. Each of the reviews contributed to measure revisions, ideally yielding a measure with improved clarity (in terms of item wording and format) that better represented the construct.

Implementation of the TK-UDA Measure

After the completion of the content validation procedures outlined, the measure was implemented. Results from the measure implementation provided evidence for the inference that performance on items accurately reflects a continuum of teacher knowledge. Specifically, evidence was garnered for the assumptions that (a) a score is a reliable estimate of levels of teacher knowledge of UDA, (b) scores for declarative and applied knowledge items differentiate types of knowledge, and (c) the measure is sufficiently broad/captures low to high levels of knowledge. The procedures, participants, and measure were the same for each of the following inquiries and are described below,

followed by the specific analytic procedures used to answer the remaining research questions.

Procedures. Participants were recruited through various personal and professional networks (emailed directly, contacted through listservs, contacted via school and district leadership). Pre- and inservice teachers were from varying geographical regions in California and Oregon. Experts from a number of states were contacted directly. Potential participants were emailed a request to participate, which included a brief description of the project and participation expectations (Appendix C). Interested participants (i.e., those who responded to the request) were sent a follow-up email that included links to the online measure (Parts 1 and 2), as well as a unique participant identification number (Appendix D). Initially, participants were asked to complete the measure within a two-week period, and reminder emails were sent at the end of the first week to encourage completion of the measure by those who had not already done so (Appendix E). From the first set of emails sent (approximately 600), 105 people indicated interest in participating. The desired number of participants was 200, so more teachers and principals were contacted and the data collection period was extended (from two weeks to eight weeks) in an effort to increase the response rate and obtain additional data. From the additional recruitment, 24 more people agreed to participate. Ultimately, 105 participants completed part 1 of the measure and 88 completed part 2. Only participants with complete data sets (i.e., parts 1 and 2) were included in the analyses ($N = 86$).

In addition, to evaluate test-retest reliability, twenty-five percent of the original sample was randomly selected and asked to complete the measure again two weeks after the conclusion of the initial implementation. Fifteen of these participants completed the

retest. Items were randomized within each of the main sections of the measure (background, declarative, and applied) to help control for threats to internal validity. Some background and demographic items were removed from the retest version, as responses were not expected to vary from the initial measure completion.

Participants. A purposive sample of experts, pre- and inservice teachers were invited to participate in the study. It was anticipated that scores from this range of participants, who differed in educational experience, would provide evidence for the measure to capture a breadth of teacher knowledge. Table 3 shows the number of participants in each subgroup and the corresponding percent of the sample.

Table 3

Participant Groups (N = 86)

Group	<i>n</i>	% of sample
Expert	4	4.7
Inservice teacher	66	76.7
Preservice teacher	16	18.6

In addition to expertise, participants were expected to range in highest degree earned, whether or not English was their primary/first language, ethnicity, and gender. This information is summarized in Table 4.

Table 4

Participant Demographics (N = 86)

	Percent by group			% Total
	Expert (n = 4)	Inservice (n = 66)	Preservice (n = 16)	
Highest degree earned				
PhD	100.0	1.5		5.8
Masters		18.2	75.0	27.9
Bachelors		80.3	25.0	66.3
Race/Ethnicity				
Hispanic		4.5		3.5
White	75.0	88.0	62.4	82.6
Asian/Pacific Islander		1.5	18.8	4.7
Bi-/Multi-Racial		1.5	18.8	4.7
Decline to state	25.0	1.5		4.7
English is primary language				
Yes	100.0	95.5	87.5	94.2
No		3.0	12.5	4.7
Decline to state		1.5		1.1
Gender				
Female	100.0	84.9	81.3	84.9
Male		12.1	18.7	12.8
Decline to state		3.0		2.3

In addition to the above demographic information, the subgroup of inservice teachers was asked questions regarding their teaching experiences. Inservice teachers varied in the number of years they had been teaching (including this year), ranging from

3 to 34 years. The average number of years taught was 14 ($SD = 8$). They also varied in grades taught (K-8), and were relatively evenly distributed across grade levels. An overall description of this participant subgroup is presented in Table 5.

Table 5

Inservice Teacher Descriptives (n = 66)

Grades taught (K – 8)	No. of participants
One grade only	41
Two grades	14
Three or more grades	11
Subjects taught	No. of participants
Elementary or single	51
Two or more	15
Subjects taught (by area)	No. of participants*
Elementary (All/Multiple)	39
Special Education	11
Language Arts	7
Mathematics	11
History/Social Sciences	1
Science/Health/Physical Education	12
Arts (Visual/Performing)/Foreign Language	9
Other (e.g., ELL/ELD, intervention)	32

Note. *Participants were asked to ‘select all that apply’; therefore, the sum of these tallies is greater than the number of participants.

Table 5 (continued)

Inservice Teacher Descriptives (n = 66)

Credentials held	No. of participants*
General education	32
Elementary/Multiple subjects	48
Secondary/Single subject(s)	17
Special education	13
Mild/Moderate disabilities	7
Moderate/Severe disabilities	3
Early childhood	6
ELL endorsement (CLAD, BCLAD, etc.)	15

Note. *Participants were asked to ‘select all that apply’; therefore, the sum of these tallies is greater than the number of participants.

The participants who completed the retest round were approximately representative of the original sample. Table 6 shows the number of participants in each of the subgroups who completed the retest round.

Table 6

Retest Participant Groups (N = 15)

Group	<i>n</i>	% retest sample
Expert	1	6.7
Inservice teacher	11	73.3
Preservice teacher	3	20.0

Measure. All participants completed the TK-UDA measure online. Data garnered from the measure implementation were analyzed using scaling, statistical, and descriptive procedures as they pertained to investigating the research questions.

Does the Measure Yield Scores That Reflect a Continuum of Teacher Knowledge?

First, correlations between scores on the background knowledge section and declarative and applied knowledge sections were calculated using PASW Statistics Grad Pack software (version 18.0; SPSS, 2010). This provided evidence for the first sub-question: Is performance on background knowledge items correlated with performance on declarative and applied knowledge items?

Next, an IRT scaling design was used to investigate whether or not scores reflected a continuum of teacher knowledge. Two additional underlying questions were necessary to support this investigation, and were addressed using a bi-factor model and IRT scaling, respectively: Are declarative and applied knowledge scores correlated, forming a single UDA knowledge measurement dimension? And, are teachers' declarative and applied knowledge of UDA scores structured from high (declarative) to low (applied)?

Analytic procedures. The research question focuses specifically on the relative difficulty of items and types of knowledge. It was hypothesized that the items and knowledge types would fall along a continuum from less (declarative knowledge) to more (applied knowledge) difficult.

Item response theory (IRT) is uniquely suited to the problem of estimating item and measurement characteristics. Whereas classical test theory conventionally characterizes item difficulty in terms of the proportion of respondents in the population

who obtain item scores with higher values (e.g., incorrect=0, correct=1), IRT conceptualizes item difficulty in terms of the amount of a trait (e.g., ability, knowledge, skill) necessary to obtain a correct response. The IRT models place item difficulty estimates on a common linear scale. For instance, equation 1 is one-parameter logistic (1PL) model, representing the most constrained of the IRT models (Embretson & Reise, 2002).

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, \dots, n \quad (1)$$

According to the 1PL model (equation 1), the probability of a correct response to a given item (P_i) is governed by the person ability parameter (θ) and the item difficulty parameter (b_i). Given the person's ability, as an item becomes more difficult, the probability of a correct response diminishes. From another perspective, given an item's difficulty, as the person's ability increases the probability of a correct response increases. Once items are estimated with the IRT model, and assuming model fit, then the relative difficulties of items and corresponding domains are available. These estimates make it possible to answer questions about relative difficulty. The IRT procedure requires estimating all items concurrently. This is possible because all respondents provided responses to all items. Using concurrent estimating, the item difficulties are properly calibrated to a common metric, and therefore comparable.

A unique feature of the current measurement system pertains to the use of scenarios for estimation of the seven UDA-specific skills. This situation creates a violation of the IRT 'local-independence' assumption. To accommodate this, the software TestFact (Wilson, Wood, and Gibbons, 2003) was used to estimate item difficulties while

taking into account the response dependencies among responses associated with any one scenario.

Using TestFact, the confirmatory bi-factor model (Gibbons and Hedeker, 1992) was estimated to test specific hypotheses that each item loads on two factors, (1) a *general* UDA factor, and (2) a *specific* factor associated with the scenario. The result of using this model is the estimation of comparable item difficulties. Furthermore, items were associated with specific trait domains (types of knowledge). The research question was whether or not these domains are differentially difficult. Once the item difficulties were estimated it became possible to test the hypotheses about the relative difficulty of the trait domains (UDA elements). A nonparametric Kruskal-Wallis test was used to test the ordering of domains (Kruskal & Wallis, 1952).

To scale all items from both the declarative and applied knowledge sections, the software Winsteps (Linacre, 2009) was used. Item difficulties for the scenario items were constrained using the parameters obtained through TestFact. This procedure provided item difficulties for the declarative (true-false) items, while anchoring the item difficulties from the scenarios (to account for the context dependency inherent in the scenario items), scaling all items along the same continuum of difficulty.

It was anticipated that the results of the IRT scaling would support the assumption that the measure captures low to high levels of knowledge and provide information regarding the structure of the measure in terms of where items and types of knowledge fell along a continuum of difficulty.

To provide evidence for test-retest reliability, correlation coefficients were calculated (using PASW Statistics Grad Pack software [version 18.0; SPSS, 2010]) for

the 15 participants for whom two sets of responses were obtained. Values greater than .6 generally indicate satisfactory to good reliability.

Does the Measure Effectively Differentiate Levels of Expertise, in Relation to (a) Teacher Knowledge of UDA (Overall) and (b) Types of Knowledge (Background, Declarative, Applied)? – Establishing Criterion-Related Evidence

Since no current measures exist against which to compare scores, observed differences in the performance of expert and non-expert groups provided evidence for evaluating discriminant validity and the assumption that the overall score was indicative of level of teacher knowledge of UDA (high vs. low), as well as differences that may be present within types of knowledge. A single-factor, non-experimental design was used to investigate this question. For each of the following analyses, PASW Statistics Grad Pack software (version 18.0; SPSS, 2010) was used.

Analytic procedures. To evaluate the significance of overall group differences, the means from the subsets of scores obtained from experts and pre-service teachers (non-experts) were compared using a Welch *t*-test (equation 2). Although attempts were made to obtain equal sample sizes for the expert and non-expert groups, these groups varied in number. The assumption of homogeneity of variance was not supported. As such, a Welch *t*-test is most appropriate.

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s\bar{x}_1^2 + s\bar{x}_2^2}} \quad (2)$$

Additionally, a multivariate analysis of variance (MANOVA) was performed to test the significance of observed differences between expert and non-expert groups using

the scores from each of the three main sections of the measure (i.e., background, declarative, applied) as outcome variables. This extension of the ANOVA method is appropriate for situations involving more than one dependent variable. Confirmation of the following assumptions supported the use of this analysis: (a) independence of observations, (b) multivariate normality, and (c) covariance among variables (Stevens, 2002). Independence of observations was assumed for the data. Although multivariate normality is difficult to characterize, “normality on each of the variables separately is a necessary, but not sufficient, condition for multivariate normality to hold” (Stevens, 2002, p. 262). In addition, this assumption can be checked, in part, by a visual analysis of scatterplots of pairs of variables, which should be, and were, approximately elliptical. Box’s test can be used to test the third assumption, homogeneity of covariance; an insignificant result indicates homogeneity (Stevens, 2002). For the MANOVA, Wilks’ Λ was calculated to test group differences, overall, on the three main sections of the measure. Posthoc pairwise multivariate tests were used to determine which groups varied significantly, and were then followed with univariate *t*-tests to further determine which variables contributed to multivariate pairwise differences.

It was anticipated that differences in levels of proficiency (high versus low) would exist between the expert and non-expert groups both overall and within each section of the measure providing evidence to support discriminant validity (i.e., the assumption that scores on the measure are indicative of level of knowledge).

Are UDA Element Domain Scores (Sub-Scores) from Applied Knowledge (Scenario) Items Useful for Identifying Professional Development Needs?

To support a more generalized use of the measure, beyond describing teachers' knowledge of UDA, element domain scores may provide evidence for targeting professional development to each of the seven UDA principles. For example, if upheld, professional development modules can be developed and implemented to address specific needs in terms of evaluating assessments for accessibility using the seven UDA principles as criteria. Multiple methods were employed to evaluate this research question by addressing the following underlying questions: (a) Are domain scores correlated, forming a single UDA skill measurement dimension? (b) Are domains differentially difficult? And, (c) do domain scores differentiate experts from non-experts?

Analytic procedures. From the IRT scaling analysis described previously, data were sampled from the scenarios to obtain domain scores for each of the seven UDA elements. Using these domain scores, correlation coefficients were calculated to evaluate the extent to which the domain scores were correlated, forming a single UDA skill measurement dimension. In addition, once the domain scores were estimated it was possible to test the hypotheses about the relative difficulty of the UDA elements. A nonparametric Kruskal-Wallis test was used to test the ordering of domains (Kruskal & Wallis, 1952).

Next, a single-factor non-experimental multivariate design (as described previously) was employed to evaluate whether or not domain scores could differentiate experts from non-experts across multiple, related dependent variables (7 UDA elements) using the data from the scenario items. A multivariate analysis of variance (MANOVA)

was performed to test the significance of observed differences between expert and non-expert groups on each of the seven UDA elements. Independence of observations was assumed for the data. The assumption of multivariate normality was evaluated, as described previously, by observing the normality of each variable and visually analyzing scatterplots of pairs of variables, which should be, and were, approximately elliptical. Box's test was used to test the assumption of homogeneity of covariance; an insignificant result indicates homogeneity (Stevens, 2002). For the MANOVA, Wilks' Λ was calculated to test group differences, overall, on the seven UDA elements. Posthoc pairwise multivariate tests were used to determine which groups vary significantly, and were then followed with univariate t -tests to further determine which variables contributed to multivariate pairwise differences.

The results section that follows describes the outcomes of each of the procedures for analyzing evidences in relation to specific research questions that align with the assumptions presented in the validity framework (described in detail in Chapter 2).

CHAPTER IV

RESULTS

Guided by the assumptions and hypotheses of the validity framework presented in Chapter II, this study sought to garner evidence to support the use of the TK-UDA measure to describe teacher knowledge of assessment accessibility issues using the principles of UDA as criteria. The following results are presented according to the evidence needed to support each research question.

Evidence of Content Validity

To garner evidence of content validity, a series of reviews were conducted. First, researchers at the University of Oregon provided an internal review. No changes were suggested for the measure blueprint or measure specifications. Clarifications were suggested for some of the items. In general, these included quantifying or providing time-delimited response categories for items regarding experience (e.g., “In the past 5 years, I have participated in training related to...”), providing descriptions for response categories that were more clear and discrete (e.g., “mostly” instead of “fairly”), and suggesting revisions to clarify directions and items. (Reviewer comments are presented in Appendices K and M). Results from this review yielded changes to the TK-UDA measure as well as subsequent review forms (i.e., the addition of more explicit review questions, as well as the existing fields for reviewer comments).

For the external and teacher reviews, three researchers with interests and research experiences in assessment and/or Universal Design in education and three teachers were asked to review the measure to further ensure clarity and consistent interpretation of items. The external/expert reviewers suggested no changes for the measure blueprint and measure specifications. Data from these reviews of the measure were aggregated. Table 7 includes tallies of responses for each of the review questions (by section).

Table 7

External and Teacher Review: Aggregated Tallies (n = 6)

Section	Are the directions clear and understandable?		Are the items clear and understandable?		Does the scale/list represent an appropriate range of responses?	
	Yes	No	Yes	No	Yes	No
1	4	0	4	1	2	2
2	6	0	4	2	6	0
3	5	1	5	1	6	0
4	5	0	5	1	6	0
5	6	0	5	1	5	1
6	5	1	4	2	6	0
7	6	0	4	2	6	0
8	5	1	3	2	4	1
9	5	0	3	2	4	1
10	5	0	1	4	5	1
11	6	0	5	1	5	1

For each 'No' tallied above, reviewers provided comments and/or suggestions to improve the directions, items, or response scale, and noted potential misinterpretations. Each of these suggestions was incorporated into measure improvements. For example, in one section, a reviewer noted, "This was a really abrupt shift from the previous items--I know that you don't want to impact your responses by too much additional information, but a brief intro like 'The following items will ask you to respond to information about testing', or something like that to help the shift." This comment led to the addition of more explicit introductions and directions for each of the major sections of the measure. Comments such as, "it may be useful on the English Language Learners question to parenthetically write English is not native or primary language" and "I have trouble with 'clearly defined constructs...' not sure what you're asking" led to revisions to improve the clarity of items. Suggestions to improve the language of the response scales included comments such as, "Could you change 'a little' to 'very little' - that might eliminate some of the potential overlap between 'a little' and 'some'?" and "Could you say 'Not at all accessible', etc. instead of just 'Not at all'?" Reviewers also provided comments (but not ratings) for the demographic items. These comments and suggestions led to revisions to response choices and language of some of the demographic items. Appendix P includes a table of all reviewer comments.

Evidence that the Measure Yields Scores that Reflect a Continuum of Teacher Knowledge

The first section of the measure, background knowledge, was comprised of items intended to measure the participants' familiarity with federal acts and regulations related to accessibility and concepts of Universal Design, their experiences working with

students of various abilities and backgrounds, and uses of technology. Inservice teachers were also asked about their experiences providing accommodations to students. Mean scores and standard deviations for each set of background knowledge items by group are presented in Table 8.

Table 8

Descriptive Statistics for Background Knowledge Items

	Group					
	Experts (<i>n</i> = 4)		Inservice teacher (<i>n</i> = 66)		Preservice teacher (<i>n</i> = 16)	
	\bar{X}	<i>SD</i>	\bar{X}	<i>SD</i>	\bar{X}	<i>SD</i>
Familiarity w/ accessibility-related regulations and Universal Design (<i>n</i> = 6)	3.00	1.00	2.54	1.02	2.82	0.91
Experiences teaching and participation in training related to teaching students of various abilities and backgrounds (<i>n</i> = 10)	2.00	1.00	1.71	0.78	2.75	1.14
Use of technology in instruction ^a (<i>n</i> = 1)	3.00	2.00	3.61	0.76	3.27	1.02
Provisions for accommodations ^b (<i>n</i> = 10)			3.18	0.90		
Total Background ^c (<i>n</i> = 27)	2.81	.90	2.66	1.08	2.84	0.99
Range (Overall mean scores)	2.18 – 3.44		1.96 - 3.56		2.13 – 4.00	

Note. ^aFollow-up items probed for personal and student uses of technology. ^bItems specific to inservice teachers. ^cTotal background does not include follow-up questions (i.e., types of training and specific uses of technology).

The first sub-question posed for this investigation regarded the correlation between performance on background knowledge items and declarative and applied knowledge items. To investigate this, PASW Statistics Grad Pack software (version 18.0; SPSS, 2010) was used to calculate correlation coefficients for the following: background and declarative, background and applied, and background and declarative + applied. Because the number of background items varied per group (preservice teachers and experts, $n = 17$; inservice teachers, $n = 27$) correlations were calculated using percent correct (percent of total possible) for each section. Table 9 shows descriptive statistics for each section, based on percent correct scoring.

Table 9

Descriptive Statistics for Each Section, Based on Percent Correct (N = 86)

Section	\bar{X}	SD	Range (%)
Background	67.36	.10	49 - 93
Declarative	76.28	.09	45 - 95
Applied	54.46	.09	33 - 74
Total	61.50	.07	44 - 74

Presented in Table 10 are the correlation coefficients. All correlations were negative, indicating inverse relationships between background knowledge and each section of the

measure, and no significant correlations were found. In addition, the correlation between the IRT scale score and performance on background items was calculated. This correlation was also negative and significant at $p < .10$.

Table 10

Correlations between Performance on Background Knowledge Items and Declarative and Applied Knowledge Items (N = 86)

		Declarative	Applied	Declarative + Applied	IRT scale score (Declarative + Applied)
Background	Pearson correlation	-.067	-.138	-.158	-.185
	Sig. (2-tailed)	.543	.205	.146	.091

The next sub-question for this investigation was whether or not declarative and applied knowledge scores were correlated, forming a single UDA dimension. Correlation coefficients were calculated using raw scores for each section. For the declarative knowledge items, the mean raw score was 15.26 (20 points possible; $SD = 1.80$); for the applied knowledge items, the mean raw score was 22.87 (42 points possible; $SD = 3.92$). The resulting Pearson Correlation Coefficient was $-.082$, $p = .452$. Correlation coefficients were also calculated using IRT scale scores, resulting in a correlation coefficient of $-.095$, $p = .690$. (See Table 11 for descriptive statistics for IRT scale scores). Thus, scores from these two parts of the measure were not significantly

correlated, indicating that perhaps the background knowledge section is assessing a different underlying construct than the other sections of the measure.

The last sub-question pertained to the structuring of participants' declarative and applied knowledge of UDA scores from high (declarative) to low (applied). As noted in the previous section, IRT scale scores for the scenario (applied knowledge) items were obtained first, then constrained and scaled with the true-false (declarative) items.

Descriptive statistics from the IRT analysis are presented in Table 11.

Table 11

Descriptive Statistics from IRT Analyses (Based on Scale Scores)

Item	\bar{X}	<i>SD</i>	Range
Declarative (<i>n</i> = 20)	-1.956	2.09	-5.61 – 1.39
Applied (<i>n</i> = 42)	-0.084	.818	-1.27 – 1.82
Total (<i>n</i> = 62)	-0.688	1.61	-5.61 – 1.82
Person	\bar{X}	<i>SD</i>	Range
Total (<i>n</i> = 86)	0.112	.362	-0.68 - 0.94

Most of the 62 items used for the IRT analysis appear to fit the model well based on mean-square residual fit statistics. Average item fit was .87 (*SD* = .17; range .40 –

1.11). Fit for two of the declarative items could not be calculated because 100% of the participants responded correctly; otherwise, items generally appear to fit the measure adequately. Four of the five items with the lowest fit (.40 - .48) were scenario items for which ‘not applicable’ was the correct response, suggesting that perhaps this response option was not useful. In terms of person fit, the average was .87 ($SD = .32$; range .52 – 2.51). In general, most of the participants’ abilities appear to be estimated adequately. However, in particular, three participants’ skill levels are over-estimated indicating a misfit between their estimated ability level and the overall pattern of person ability estimates.

Figure 3 shows the distribution of participant scale scores on the left of the midline, from more ‘ability’ (i.e., knowledge of UDA; top) to less (bottom). Person scores ranged from -.68 to .94 ($\bar{X} = .112$, $SD = .362$). On the right of the midline, the distribution of items is presented from most difficult (top) to least difficult (bottom). Item difficulties ranged from -5.61 – 1.82 ($\bar{X} = -.68$, $SD = 1.61$). Items in bold are true-false (declarative), labeled ‘tf’ with an item number (e.g., **tf1 – tf120**); items in italics are scenario (applied), labeled with the scenario and item/UDA element number (e.g., *s1i1 – s6i7*). As shown in the item distribution presented in Figure 3, items in the applied knowledge section were generally more difficult than items in the declarative knowledge section, with the exception of a few true-false items that fell toward the more difficult end of the scale.

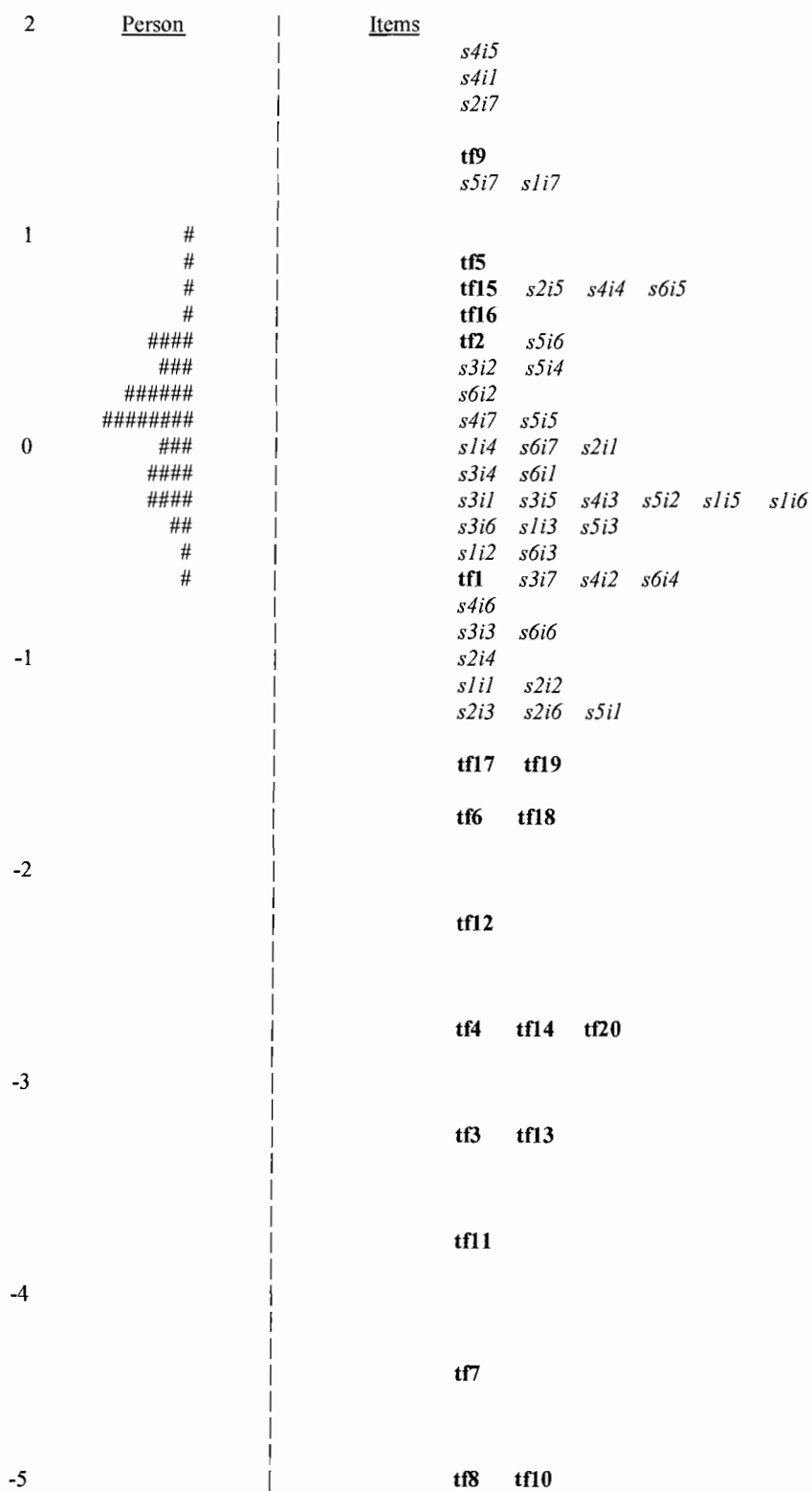


Figure 3. Distribution of person scores and item difficulties. (Note: Each 'x' represents 2 participants).

Each of the items in the declarative and applied knowledge sections is related to one of the seven UDA elements (as noted in the Methods chapter). Table 12 shows the distribution of items across elements. Item labels are consistent with those presented in Figure 3.

Table 12

Items from Declarative and Applied Knowledge Sections Pertaining to Each UDA Element

UDA element	Items per section	
	Declarative (true-false)	Applied (scenario)
Simple, clear, and intuitive instructions and procedures	tf14, tf19, tf20	s1-s6i1
Maximum readability and comprehensibility	tf5, tf11, tf16	s1-s6i2
Maximum legibility	tf12, tf15, tf17	s1-s6i3
Inclusive assessment population	tf1, tf2, tf6	s1-s6i4
Precisely defined constructs	tf7, tf9, tf13	s1-s6i5
Accessible, non-biased items	tf3, tf10	s1-s6i6
Amenable to accommodations	tf4, tf8, tf18	s1-s6i7

In addition to the selected response items used for the previous analyses, participants completed two constructed response scenario items. These items represented a range of UDA elements and extended the application of UDA to address the role of the teacher as ‘retrofitter’. Participants were asked to provide suggestions for improving the setting, directions, and test items within each scenario to improve its accessibility. Although responses were coded by UDA element, because participants were not asked to specifically comment on accessibility with regard each UDA element, this data was not aggregated with that used in the previous analyses. Table 13 shows descriptive statistics for these two items based on the overall number of suggestions for improvement coded by each UDA Element (7 points possible).

Table 13

Descriptive Statistics for Constructed Response Scenario Items

	<i>n</i>	\bar{X}	<i>SD</i>	Range
Scenario 1	76	2.47	1.16	0 - 6
Scenario 2	73	3.05	1.29	0 - 6

Presented in Tables 14 and 15 are examples of comments and suggestions provided by participants for each of the constructed response scenario items as they relate to each of the seven UDA elements. In general, participants either commented on part(s) of the scenario that might present a barrier for students in terms of accessibility, they provided a suggestion for improving the scenario, or they rewrote a section of the scenario to improve its accessibility. No comments were provided for either scenario

pertaining to ‘inclusive assessment population’; therefore it is not represented in the summary tables.

Table 14

Scenario 1: Examples of Constructed Responses (Comments and Suggestions for Improving Accessibility) Coded by UDA Element (N = 76)

UDA element	<i>n</i>	Comments/Suggestions for improving accessibility
Simple, clear, and intuitive instructions and procedures	57	<p>“Students might not know what a scantron is. If there's going to be a 'bubble' it should be called an answer choice and there should be an illustration of the expectations for filling it in”</p> <p>“The directions are confusing. Students should be told first to find the answer to the question using the test to write on. Then bubble in the corresponding or correct answer.”</p> <p>“You are going to try to solve a riddle. Read the entire riddle and think about what information you have, and what information you do NOT have. Solve each part of the riddle. You may write your answer under each item, or on scrap paper, but make sure you show your work. You may use a calculator if you like, but make sure you show your work!”</p>
Maximum readability and comprehensibility	44	<p>“Too confusing to understand the multi-step instruction...allowing the student to provide an answer after each step would help a teacher understand where they went wrong in deriving their answer.”</p> <p>“The test problem is very confusing. The riddle's directions do not make sense. For example, it says there is a one in the thousands place, but the answers do not have a one in that place.”</p> <p>“Be clearer in the items leading to the answer.”</p> <p>“Clues are very confusing. For example: 'Multiply the digit in the thousands place by 2.' And then do what with it?”</p>
Maximum legibility	20	<p>“Spacing of clues. The page looks to cluttered... More spacing between scenarios.”</p> <p>“Bullet points for directions instead of a paragraph. Use an icon with a slash for no calculators.”</p> <p>“Bold the directions so they stand out from the test items.”</p>

Table 14 (continued)

Scenario 1: Examples of Constructed Responses (Comments and Suggestions for Improving Accessibility) Coded by UDA Element (N = 76)

UDA element	<i>n</i>	Comments/Suggestions for improving accessibility
Precisely defined constructs	10	<p>“I have no idea what this is intended to measure - this needs to be made more clear”</p> <p>“The item is assessing multiple skills- reading, math, logic, contextualization, etc.”</p> <p>“Is this about math? It seems to be more about following convoluted directions.”</p> <p>“It's unclear what concept this question measures: logic, calculation; following multi-step directions. Break the question into multiple, clearer questions focused on a single concept.”</p>
Accessible, non-biased items	20	<p>“Remove the Batman and Riddler passage and replace it with a more direct question. It is culturally bias and convoluted”</p> <p>“Remove the references to the Riddler and Batman as many students may have no experience with these characters and may make incorrect inferences in their answers based on the unknown information in the test item.”</p> <p>“I would also change clue four to something with numbers/math versus language comprehension For example, I might use '...is the number of sides on a square,' because ELL students may not understand 'a hand without a thumb.’”</p>
Amenable to accommodations	37	<p>“The test administrator may record the student's responses for each segment of the test item. The administrator may read the student's responses for the student to evaluate the correct multiple choice answer.”</p> <p>“I might allow the use of a calculator as an accommodation for some of the students.”</p> <p>“I would allow students to have the test read to them because it is only testing their math skills not reading ability. If a student needs accommodation of using a calculator I would allow its use. Finally, I would allow a scribe to transfer answers onto the scantron for students who may have difficulty transferring their answers.”</p>

Table 15

Scenario 2: Examples of Constructed Responses (Comments and Suggestions for Improving Accessibility) Coded by UDA Element (N = 73)

UDA element	<i>n</i>	Comments/Suggestions for improving accessibility
Simple, clear, and intuitive instructions and procedures	55	<p>“The directions need to be more specific, such as the above paragraph (as there may be one below). Also, the 'if needed' should be more specific on when to give evidence.”</p> <p>“Read the selection. Answer the questions using evidence to support answers.”</p> <p>“Ten minutes to respond to a test in English? Really? Sounds like a very stressed, rushed environment for students trying to learn an alternate language.”</p> <p>“Students should have more time to read the passage and answer each question about the passage. They should be given at least 30 minutes.”</p>
Maximum readability and comprehensibility	47	<p>“Very poorly written paragraph - informal style that is distracting and does not correspond with the test questions.”</p> <p>“The sentences are long and filled with difficult language.”</p> <p>“The topic should be something the 7th grade students can relate to that include vocabulary they have previously learned. Questions should relate directly to the passage.”</p> <p>“I would change the content to an accessible, age appropriate, subject for all students.”</p>
Maximum legibility	61	<p>“Font is hard to read, needs to be sans serif.”</p> <p>“Items should be retyped in a different font. Lines should be longer for answers.”</p> <p>“The problem should be printed, not in cursive writing, as not all students can read cursive.”</p> <p>“Change the font type, size, and line spacing.”</p>

Table 15 (continued)

Scenario 2: Examples of Constructed Responses (Comments and Suggestions for Improving Accessibility) Coded by UDA Element (N = 73)

UDA element	<i>n</i>	Comments/Suggestions for improving accessibility
Precisely defined constructs	14	<p>“The first thing I would do is look at the test item specifications and to understand what content the item was testing.”</p> <p>“...focus is on math, not comprehension.”</p> <p>“Test is meant to measure comprehension, not ability to build an argument; therefore, questions should not require students to provide evidence”</p>
Accessible, non-biased items	25	<p>“Topic is biased, and should be changed. Assumes kids know about credit cards and budgets.”</p> <p>“The vocabulary is quite advanced for ELL. Many 7th graders may not have experienced 'bounced checks,' budget, and may not have a clue about minimum payments on credit cards. They must have those experiences first.”</p> <p>“Vocabulary usage is also difficult for ELLs, with such phrases as: 'carried a balance' (picked it up and carried it where?), debt spiraling out of control (literally?), bounced checks (how high do they bounce?), etc.”</p>
Amenable to accommodations	21	<p>“Students should be allowed to use a dictionary, and have extended time as needed.”</p> <p>“Translator should be made available for ELL students.”</p> <p>“...the setting may need to be changed for students that need additional help. In a quiet environment with one on one support, if needed.”</p>

Evidence of Reliability

Evidence for reliability, in general, supports both the internal consistency and test-retest reliability of the measure. From the IRT analysis, a strong item reliability of .94 was obtained, indicating that the declarative and applied knowledge items represent a range of difficulty. A relatively weak person reliability of .28 most likely indicates that

the sample of participants did not represent a wide range of abilities. In addition, this could also be due to an insufficient number of items for the declarative and applied knowledge sections, or, since the items were dichotomously scored, may have been a result of the number of response categories per item (Linacre, 2009). In addition, Cronbach's alpha was calculated for the declarative and applied knowledge sections ($\alpha = .248$ and $.827$, respectively), as well as the total (declarative + applied; $\alpha = .781$). This measure of reliability indicates that the true-false items may not be assessing the same construct (declarative knowledge of UDA), whereas the scenario items appear to be measuring relatively the same construct

To provide evidence for test-retest reliability, correlation coefficients were calculated for the 15 participants for whom two sets of responses were obtained. Table 16 provides descriptive statistics, including means and standard deviations, for Times 1 and 2 for each main section of the measure and total. For the background knowledge items, participant responses were based on a 4-point scale (24 points possible). For the declarative and applied knowledge sections, responses to each item were scored correct (1) and incorrect (0). Table 17 shows the Pearson Correlation Coefficients between times 1 and 2 per section and overall (total) score.

Table 16

Descriptive Statistics for Test-Retest (N = 15)

Section	Time 1		Time 2	
	\bar{X}	<i>SD</i>	\bar{X}	<i>SD</i>
Background (n = 6; 4-point scale)	14.93	5.24	14.47	4.36
Declarative (n = 20)	15.40	1.77	16.40	1.72
Applied (n = 42)	24.07	3.28	22.93	2.96
Total (n = 68)	54.40	6.57	53.80	4.60

Table 17

Test-Retest Correlations

Time 2	Time 1			
	Background	Declarative	Applied	Total
Background	.824**			
Declarative		.484*		
Applied			.536**	
Total				.636**

Note. * $p < .10$; ** $p < .05$

In general, correlations were moderate (.484) to high (.824), indicating satisfactory test-retest reliability.

Criterion-Related Evidence

To obtain evidence for criterion-related validity, differences between groups based on level of expertise were evaluated. First, a Welch *t*-test was conducted using total scores from expert and preservice teacher groups to see if differences existed between the two extremes on overall scores. No significant difference was found between expert and preservice groups ($t(18) = 1.152, p = .264$). (In Table 18, below, descriptive statistics are presented per group for each section of the measure). Follow-up *t*-tests were conducted to explore whether significant differences existed between these two groups on any of the three main sections of the measure. A significant difference between groups was present only for the declarative knowledge section ($t(18) = 2.149, p < .05$).

Next, a MANOVA was conducted to evaluate whether the measure effectively differentiates levels of expertise in relation to types of knowledge (background, declarative, and applied). Table 18 contains descriptive statistics by group for each section of the measure. For this analysis, independence of observations was assumed. Boxplots and histograms were examined for univariate and multivariate normality, and no section subscales presented significant deviations from normality. Two outliers were present on the declarative knowledge subscale. No significant mean differences resulted after removing the outliers; therefore, they were not considered influential.

Table 18

Descriptive Statistics Per Group for Each Section, Based on Percent Correct (N = 86)

Section	Expert		Preservice		Inservice	
	\bar{X}	<i>SD</i>	\bar{X}	<i>SD</i>	\bar{X}	<i>SD</i>
Background	68.75	.11	66.50	.10	70.59	.11
Declarative	88.75	.05	75.30	.08	77.19	.10
Applied	52.97	.10	55.12	.09	52.08	.10
Total (Overall)	67.36	.10	76.28	.09	54.46	.09

Homogeneity of variance was tested using Box's Test of Equality of Covariance Matrices and Levene's Test of Equality of Error Variances. Examination of Box's M revealed heterogeneity of variance, indicating that the observed covariance matrices of the dependent variables (section subscale) do differ significantly across groups ($F(12, 287) = 1.806, p = .047$). Levene's Test of Equality of Error Variances yielded non-significant results for all section subscales, indicating error variance of the dependent variables does not differ to a significant degree across groups. The between-subjects multivariate results indicated a statistically significant difference in the multivariate combination of the section subscores based on level of expertise, Wilks' $\Lambda = .863, F(6, 162) = 2.063, p < .10$. Results of the univariate tests indicated a statistically significant difference based on level of expertise for the declarative knowledge section ($p \leq .05$) only.

In addition to the MANOVA, a non-parametric Kruskal-Wallis rank order test was used to analyze differences between groups based on IRT person scale scores. Because the Kruskal-Wallis is non-parametric, it is a more sensitive test of group differences. Results from this analysis indicate no significant differences between groups based on IRT scale score (chi-square $[2, N = 3] = 1.501; p = .472$).

Evidence Supporting the Use of the Measure for Identifying Professional Development Needs

First, correlations between domain scores were calculated using item difficulties for each domain (i.e., UDA element within the applied knowledge section). These were evaluated to determine if the domains formed a single UDA skill measurement dimension. Three pairs of domains had significant correlations ($p < .10$):

- ‘Simple, Clear, and Intuitive Directions and Procedures’ and ‘Precisely Defined Constructs’ (Pearson Correlation Coefficient = .894, $p < .05$);
- ‘Maximum Legibility’ and ‘Inclusive Assessment Population’ (Pearson Correlation Coefficient = .818, $p < .05$);
- ‘Maximum Readability and Comprehensibility’ and ‘Amenable to Accommodations’ (Pearson Correlation Coefficient = .772, $p < .10$).

‘Accessible, non-biased items’ was not significantly correlated with any other element.

These correlations indicate that the measure is assessing different skill dimensions within the applied knowledge section. A correlation matrix is presented in Table 19.

Table 19

Correlation Matrix for UDA Element Domain Scores

UDA Element	UDA Element						\bar{X}	SD
	1	2	3	4	5	6		
1							-.167	1.10
2	-.226						-.298	.555
3	.161	.170					-.608	.362
4	.267	.116	.818**				-.095	.641
5	.894**	-.426	.253	.213			.482	.620
6	-.582	.368	.471	.603	-.559		-.513	.924
7	-.428	-.772*	-.169	-.234	-.121	.065	.612	.818

Note. * $p < .10$; ** $p < .05$; (1) Simple, clear, and intuitive instructions and procedures; (2) Maximum readability and comprehensibility; (3) Maximum legibility; (4) Inclusive assessment population; (5) Precisely defined constructs, (6) Accessible, non-biased items; (7) Amenable to accommodations.

The next question for this investigation was whether or not domain scores (subscores) from the applied knowledge items were useful for identifying professional development needs. First, item difficulties were sampled from the IRT scaling of scenario items. Next, using the Kruskal-Wallis rank order test, items were ranked in order of difficulty and the mean rank per element was calculated. Table 19 contains the mean

element difficulties and standards deviations, as well as the mean rank per element. As indicated by the chi-square statistic, $\chi^2(6, N = 7) = 12.373, p \leq .05$, the UDA elements appear to be differentially difficult. Therefore, based on the applied knowledge items, professional development needs could be targeted at the domain level.

Table 20

Kruskal-Wallis Test of UDA Element Scale Scores (Sorted by Mean Rank from Least to Most Difficult)

UDA element ($n = 6$ per element)	Element difficulty		Mean rank
	\bar{X}	SD	
Maximum legibility	-.608	.362	13.00
Accessible, non-biased items	-.513	.924	14.42
Maximum readability and comprehensibility	-.298	.555	18.92
Simple, clear, and intuitive instructions and procedures	-.167	1.10	19.25
Inclusive assessment population	-.095	.641	23.33
Precisely defined constructs	.482	.620	30.75
Amenable to accommodations	.612	.818	30.83

Note. Chi-Square ($6, N = 7$) = 12.373, $p \leq .05$

The final question for this investigation was whether or not domain scores differentiated experts from non-experts. Table 20 contains descriptive statistics by group

for each UDA element (domain). To evaluate this question, a MANOVA was conducted, with level of expertise as the grouping variable. Independence of observations was assumed. Boxplots and histograms were examined for univariate and multivariate normality, and no element subscales presented significant deviations from normality. A few outliers were present on four of the seven element subscales. No significant mean differences resulted after removing outliers; therefore, outliers were not considered influential. Homogeneity of variance was tested using Box's Test of Equality of Covariance Matrices and Levene's Test of Equality of Error Variances. Examination of Box's M shows there was homogeneity of variance, upholding the assumption that the observed covariance matrices of the dependent variables (element subscale) do not differ significantly across groups ($F(28, 2611) = 1.020, p = .436$). Levene's Test of Equality of Error Variances yielded non-significant results for all but one element subscale, indicating error variance of the dependent variables does not differ to a significant degree across groups, with the exception of the legibility subscale.

The between-subjects multivariate results indicated a non-statistically significant difference in the multivariate combination of the UDA element subscale scores based on level of expertise, Wilks' $\Lambda = .809, F(14, 154) = 1.230, p > .05$. Results of the univariate tests indicated a statistically significant difference based on level of expertise for the legibility subscale only ($p = .039$), consistent with the results of the tests of homogeneity of variance previously reported. Based on these results, domain scores do not differentiate levels of expertise.

Table 21

Descriptive Statistics by Group for Each UDA Element (Based on Raw Score Per Element)

	Expert (<i>n</i> = 4)	Inservice (<i>n</i> = 66)	Preservice (<i>n</i> = 16)	Total (<i>N</i> = 86)
UDA element	\bar{X} (<i>SD</i>)	\bar{X} (<i>SD</i>)	\bar{X} (<i>SD</i>)	\bar{X} (<i>SD</i>)
Simple, clear, and intuitive instructions and procedures	4.00 (1.41)	3.52 (1.03)	3.38 (1.26)	3.51 (1.08)
Maximum readability and comprehensibility	3.00 (1.16)	3.65 (1.20)	3.81 (1.28)	3.65 (1.21)
Maximum legibility	5.00 (.816)	4.45 (1.23)	3.56 (1.86)	4.31 (1.38)
Inclusive assessment population	3.25 (1.26)	3.24 (1.30)	3.06 (1.06)	3.21 (1.25)
Precisely defined constructs	1.25 (1.50)	2.21 (1.34)	2.31 (1.30)	2.19 (1.34)
Accessible, non-biased items	3.50 (1.00)	4.15 (1.10)	3.75 (.931)	4.05 (1.07)
Amenable to accommodations	2.25 (.500)	1.92 (.933)	2.00 (.816)	1.95 (.893)

In the next chapter, a discussion of these results is presented as they relate to the evidence needed to support the validity argument. In addition, limitations of the study and directions for future research are addressed.

CHAPTER V

DISCUSSION

The purpose of this dissertation research was to design and validate a measure of teacher knowledge of Universal Design for Assessment (UDA). The measure was designed to capture teacher knowledge along a continuum from background to declarative to applied through a variety of item and response types. The UDA elements presented by Thompson, Johnstone, & Thurlow (2002), provided a framework and criteria for teachers' evaluation of assessment accessibility issues. The importance of teacher knowledge in this area is reinforced by the diversity of students in today's classrooms, efforts to increase access to and inclusion in general education curricula and accountability assessments, and efforts to close the achievement gap between educationally disadvantaged students and their peers. By improving accessibility, through appropriate accommodations and applications of UDA, a wider range of students can effectively participate in learning and evaluation, and the interpretations of student performance that contribute to the instructional decision-making process can be made with greater confidence and accuracy.

In the following sections, the results presented in the previous chapter are summarized and interpreted as they pertain to the evidence needed to support the validity

framework. Implications and considerations for measure revisions are discussed throughout. In addition, limitations and directions for future research are addressed.

Evidence of Content Validity

The first assumption of the validity argument stated that the seven UDA elements were inclusive of all UDA principles. Evidence for this assumption came from a review of the literature and provides initial support for content validity. The content of the TK-UDA was based upon and derived from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999). Items were designed to capture a range of knowledge. Background knowledge items were intended to capture information about participants' familiarity with regulations related to accessibility and universal design concepts, their experiences and training related to teaching students of varying abilities and backgrounds, use of technology in education, and provisions for accommodations. Declarative knowledge items represented a range of UDA elements and were designed to measure factual knowledge. Applied knowledge items were contextualized within scenarios representing positive and negative examples of each UDA element.

The second assumption stated that the measure (TK-UDA) was representative of the seven UDA elements. Results from each of the reviews (internal, expert, and teacher) provide evidence for this assumption and additional support for the content validity of the measure. In general, the measure reviews yielded changes that improved the content and clarity of the measure. Although solicited, no changes were suggested by any of the reviewers for the measure blueprint or specifications. Overall, changes suggested for

improving the measure were not related specifically to content representativeness; rather, they were associated with the language and clarity of the directions, items, and response scales. This may indicate that the review questions were not specific enough to capture this information or, alternatively, that the content of the measure was sufficiently representative of the UDA elements and construct, overall. Additional measure reviews would provide clarification for this interpretation.

Evidence that the Measure Yields Scores that Reflect a Continuum of Teacher Knowledge

Three assumptions within the validity argument were presented to guide the collection of evidence to support the claim that the measure yields scores that reflect a continuum of teacher knowledge. Evidence was first obtained to evaluate the correlation between performance on background knowledge items and declarative and applied knowledge items. All correlations were negative, indicating inverse relationships between these sections of the measure. Based on correlations calculated using percent correct scores, none of the correlations were significant. However, the correlation between background knowledge items and IRT scale score was significant ($p < .10$), indicating an inverse relationship between background knowledge and estimated ability level. This may be due to the bifactor analysis accounting for the context and difficulty of the scenario items, and subsequent IRT scaling (whereas the percent correct scores were based on raw data, not the relative item difficulty).

Next, evidence was garnered to evaluate whether or not declarative and applied knowledge sub-test scores were correlated, forming a single UDA measurement dimension. Again, using percent correct scores, then IRT scale scores, correlation

coefficients were calculated. Both yielded non-significant correlations between these two sections of the measure. This indicates that perhaps these sections are assessing different underlying constructs, for example, ‘types of knowledge’ (declarative and applied) as opposed to one underlying UDA construct.

Last, evidence was obtained to support the structuring of items and scores from high (declarative) to low (applied). In general, as hypothesized, items within the declarative knowledge section were less difficult than items in the applied knowledge section. Although a strong item reliability value was obtained, indicating that declarative and applied knowledge items represent a range of difficulty, when considering the placement of items along the continuum by knowledge type, overall, declarative items represented a rather narrow range of difficulty and, with the exception of a few items, were primarily ‘easy’. Applied (scenario) items, though relatively more difficult in comparison to declarative items, also represented a narrow range of difficulty. In addition, four of the five the items with the poorest fit were scenario items for which ‘not applicable’ was the correct response. (For five of the scenario items, this was the correct response). Person scores also represented a rather narrow range of ability, indicated by both the range of scores, as well as the relatively weak person reliability of .28.

Given the results of the item scaling, revisions to the measure might include adding items of varying difficulty, as well as revisions to existing items, to increase and improve the range of item difficulty. To address the misfit of the items with ‘not applicable’ as the correct response, considerations for measure revision include either eliminating this as a response option and revising the scenarios to represent only positive and negative examples of each UDA element, or perhaps using a Likert-type scale for

evaluation of accessibility based on each UDA element (which would require analytic techniques that permit polytomous scoring). The weak person reliability index, could be addressed by obtaining a greater sample size, which may represent a wider range of abilities, adding more items to the declarative and applied knowledge sections, or, as noted above, revising the response scale to include more categories.

The two constructed response items provided additional, qualitative information that revealed participants' abilities to evaluate the test scenarios for accessibility. In addition, these items captured information related to teachers' role as test retrofitter. Although participants were not asked to specifically provide suggestions to improve accessibility based on each UDA element, their comments were coded and generally fell into one of the seven UDA element categories. The most comments were provided for the UDA elements of 'simple, clear, and intuitive directions and procedures', 'maximum readability and comprehensibility', and 'maximum legibility', perhaps because these elements were easier to evaluate (or more apparent) within the given contexts or easy to identify visually. Fewer overall comments were made related to the UDA elements of 'precisely defined constructs', 'accessible, non-biased items', and 'amenable to accommodations', perhaps because these may be more difficult to evaluate. No comments related to 'inclusive assessment population' were given for either constructed response scenario. However, given the context of the second scenario ("ELL students in Mrs. Angeli's 7th grade class are given brief weekly reading comprehension assessments"), it was expected that this might have been considered a non-inclusive population. One consideration for revising the constructed response items might include changing the response prompts from 'setting', 'directions', and 'test item' to the seven

UDA elements. In addition, although these items provided information regarding participants' abilities to evaluate the scenarios for accessibility issues, they are somewhat contrived. Another consideration for revising the constructed response items would be to have teachers evaluate and revise actual student assessments.

Evidence of Reliability

Reliability evidence, in general, supported both the internal consistency and test-retest reliability of the measure. Reliability was assessed using Cronbach's α and test-retest correlations, respectively. (Reliability estimates from the IRT analysis were discussed previously). Cronbach's α was used to evaluate the degree to which the items on the measure combined to assess a single trait. The overall reliability (for declarative and applied knowledge sections) was moderately strong, $\alpha = .78$. For the declarative knowledge items, reliability was weak, $\alpha = .248$, indicating that the true-false items are measuring other factors not captured by the measure, in addition to 'declarative knowledge of UDA'. For the applied knowledge items, $\alpha = .827$, indicating that the scenario items are reliably assessing 'applied knowledge of UDA'.

To provide evidence for test-retest reliability, correlation coefficients were calculated for the 15 participants for whom two sets of responses were obtained. Correlations between times 1 and 2 were moderate to strong for the background and applied knowledge sections, as well as the total, (.824, .536, and .636, respectively). The correlation between times 1 and 2 for the declarative knowledge section was weaker than the other sections, .484, indicating that performance on this section was not consistent across administrations, and perhaps, a practice effect for this section of the measure.

Criterion-Related Evidence

To support the assumption that scores on the measure were indicative of levels of teacher knowledge of UDA, evidence for group differences was evaluated. Since there are no existing measures of similar constructs against which relevant criterion might be evaluated, differences in group performance on the measure, or discriminant validity, were used to test the hypothesis that scores were indicative of different levels of proficiency. Differences between groups were evaluated a number of ways. First, a Welch *t*-test was conducted using total scores from expert and preservice teacher groups to see if differences existed between the two extremes on overall scores. No significant difference was found between groups, $t(18) = 1.152, p = .264$. Follow-up *t*-tests were conducted to examine whether significant differences existed between experts and preservice teachers any of the three main sections of the measure. A significant difference between groups was present only for the declarative knowledge section, $t(18) = 2.149, p < .05$. Experts performed significantly better than preservice teachers on this section of the measure (\bar{X} percent correct = 88.75 and 77.19, respectively).

Next, a MANOVA was conducted to evaluate whether the measure effectively differentiates levels of expertise in relation to types of knowledge (background, declarative, and applied). The between-subjects multivariate results indicated a statistically significant difference based on level of expertise. Results of the univariate tests indicated a statistically significant difference based on level of expertise for the declarative knowledge section ($p \leq .05$) only, consistent with the results of the *t*-tests.

Last, a non-parametric Kruskal-Wallis rank order test was used to analyze differences between groups based on IRT person scale scores. Results from this analysis indicate no statistically significant differences between groups based on IRT scale score.

Based on the results, discriminant validity of the measure, overall, is not upheld; the measure does not effectively differentiate levels of expertise. This may have been influenced by a number of variables, including the sample size, narrow range of abilities within the sample, and narrow range of item difficulties. In addition, this may have been a result of the way in which experts were identified and defined (as researchers with experiences studying UDA or educational assessment). Given a broader sampling of expertise across groups (e.g., pre- and inservice teachers with extensive applied experience with educational assessment), this assumption may have been better supported (i.e., the measure may differentiate high versus low levels of knowledge). This hypothesis may be evaluated with additional research. Also, because ‘experts’ (who included university faculty and researchers) may be contributing to curriculum development and instruction in credential courses, preservice teachers may be exposed to universal design and/or UDA concepts in their credentialing programs, therefore these two groups could conceivably perform similarly. Although the measure appears to differentiate levels of expertise based on declarative knowledge scores, the weak internal reliability for this section of the measure suggests that more than one trait is being assessed, which, without additional analyses, makes interpreting differences in ‘declarative knowledge’ difficult.

Evidence Supporting the Use of the Measure for Identifying Professional Development Needs

The final assumption of the validity argument stated that domain scores (UDA element sub-scores) were useful for identifying professional development needs. A variety of evidences were evaluated to support this assumption. First, correlations between domain scores were calculated using item difficulties for each UDA element within the applied knowledge section. These were evaluated to determine if the domains formed a single UDA skill measurement dimension. Three pairs of domains had significant correlations ($p < .10$), and one was not correlated with any other, indicating that, within this section, more than one skill dimension is being assessed. Next, to evaluate whether or not domain scores (subscores) from the applied knowledge items were useful for identifying professional development needs, item difficulties were sampled from the IRT scaling of scenario items, items were ranked in order of difficulty, and the mean rank per element was calculated. Results from the Kruskal-Wallis rank order test indicate that the UDA elements appear to be differentially difficult, potentially signifying that professional development needs could be targeted at the domain level. Last, a MANOVA was conducted to evaluate whether or not domain scores differentiated experts from non-experts. Non-significant results were obtained, indicating that this domain scores, overall, do not differentiate levels of expertise.

Based on these results, because the domains are differentially difficult, it may be possible to target professional development at the UDA element level. However, since the domain scores do not differentiate levels of expertise, misassignment of participants to professional development modules based on UDA elements is possible. Perhaps setting

a passing score for each domain (e.g., 4 of 6 items correct) would provide a better means of comparing groups and assigning participants to professional development specific to their needs.

Consequences of Score Use and Considerations for Measure Revisions

The proposed uses of the measure for this study were to describe teachers' knowledge of universal design for assessment and to provide initial evidence for its usefulness in identifying professional development needs at the UDA element level. The validity argument outlined a chain of inferences, assumptions, and evidences intended to support the use of the measure for these purposes. However, the reliability of the validity framework relies heavily upon the information garnered at each stage of the evidentiary process; that is, the inferences and assumptions are upheld or refuted based on the results of the analyses. In general, the measure appears to (a) be representative of the seven elements of UDA, (b) capture a range of teacher knowledge and represent a range of item difficulty, and (c) be potentially useful for identifying professional development needs. Discriminant validity of the measure was not upheld, for a number of reasons, including those described previously. Although the measure appears to capture a continuum of knowledge with a range of easy and difficult items, the participant sample represented a rather narrow range of ability, and the majority of the items on the measure, especially those within the declarative knowledge section were 'easy' items.

The results of this study provide evidence that indicates the need for measure revisions before the claim can be made that the TK-UDA accurately describes levels of teacher knowledge of Universal Design for Assessment. One consideration would be to revise or simplify the language of the seven UDA elements or reduce the number of UDA

elements to be applied at the classroom level. For example, ‘clear directions’ might capture the essence of ‘simple, clear, and intuitive directions and procedures’ and would eliminate the issue of evaluating such a compound statement. In addition, it is possible that some of the elements are less applicable to classroom assessments than they are to large-scale assessments. For example, ‘inclusive assessment population’ could be eliminated as a consideration for classroom level assessments because the context of the population to be assessed is limited to the students within the class.

Other considerations for revision include: adding more difficult declarative knowledge items (or revising these items to represent a wider range of difficulty); eliminating ‘not applicable’ as a response option for the applied knowledge items and revising the scenarios to represent only positive and negative examples of each UDA element; using a Likert-type scale for evaluation of accessibility based on each UDA element within the scenarios; and adding (or replacing the scenario items with) a section that includes actual student assessment examples, that are more realistic than the existing scenario items.

After measure revisions are made, another series of reviews should be conducted that includes review items more specific to the appropriateness and representativeness of measure content/items to the underlying construct of UDA. The revised measure would then be implemented, ideally completed by a larger sample of participants. With a larger sample size, additional analytic techniques could be used, such as confirmatory factor analysis, and results could be interpreted with greater confidence.

Limitations

There are several limitations that should be considered when interpreting the results of this study. First, and most significant, is the small sample size; in addition, there are limitations related to the measure and the analyses.

Limitations of sample size. A number of factors may have contributed the small sample size. First, related to getting participation information to teachers, some school districts had established procedures and requirements for research involving district employees and students. Given the time required to complete the approval processes, in most cases at least a month, I decided not to pursue participant recruitment in these districts as the process for approval would have extended beyond my timeline for data collection. In addition, principals who were contacted may have elected not to forward information about the study and participation opportunity to their teaching staff.

Second, participants were purposefully selected based on their affiliation with one of the three target groups: expert, inservice teacher, or preservice teacher. This was done with the intention of garnering responses from a broad range of participants to examine the ability of the TK-UDA to capture a continuum of knowledge from low to high. In addition, participation was voluntary; therefore, the sample only includes participants who elected to complete the measure. Decisions regarding participation may have been affected by time, familiarity (or lack thereof) with the topic of UDA, and/or incentives for participation. In short, the people in each of the target groups are busy and have responsibilities that require time commitments beyond the hours in a school day; they may simply not have had time to participate in this study. Some may have chosen not to participate because of a lack of familiarity with accessibility issues or Universal Design

for Assessment; and, although a description was provided in the introductory letter, they may not have felt comfortable participating. The converse may also be true – those with knowledge of accessibility issues and UDA may have been more willing to participate. Another factor affecting participation may have been the incentives offered. Incentives for participation included a \$10 Amazon gift card for completing Part 1 and a drawing opportunity for one of 4 iPod shuffles for completing Part 2. Although these incentives were offered, given the time projected for completing both parts of the survey (approximately 40 minutes), the incentives may not have been enticing to some.

A third noted limitation, related to the sampling procedures used, is the lack of precision in response rate. Because information regarding participation was broadly distributed and participants were recruited through listservs and school leadership, in addition to personal and professional networks, calculating an exact response rate is not possible. However, the completion rate, that is, the percentage of participants who agreed to complete the measure and did, can be calculated. As noted in the Chapter III, 129 people agreed to participate, and 86 completed both Parts 1 and 2, yielding a completion rate of approximately 67%. Although the completion rate was moderately high, it was impacted by the measure being presented in two parts; some participants did not complete both parts, and only complete data sets were included in the analyses.

Limitations of the measure. Several limitations are related to the design of the measure. First, is the lack of flexibility in using the seven elements of UDA as criteria for evaluating accessibility. Although this point did not arise during the measure/content review, it was discussed with the dissertation committee at the time of proposal. Inherent in Universal Design is the idea that accessibility is considered within the design stages,

rather than as a retrofit to existing materials. Applying the UDA elements as criteria for evaluating the accessibility of existing assessment, much like a checklist, may be considered contradictory to the flexibility of universal design. However, as noted in the literature review, the intent of applying the UDA elements as criteria for evaluating accessibility was to ensure that tests (and other instructional materials) meet minimum or baseline requirements for accessibility. Also, because teachers use tests (and instructional materials) that may not have been designed to be accessible to all of their students, they need to be able to identify these issues in order to assign accommodations, interpret student performance, and make instructional decisions based on assessment results.

The next limitation is related to survey design, in general. By maintaining the language of the UDA elements for the applied knowledge items, participants were asked to evaluate multiple questions in one item. This pertains to three elements in particular: simple, clear and intuitive instructions and procedures (6 considerations in one), maximum readability and comprehensibility (2 considerations) and accessible, non-biased items (2 considerations). Although efforts were made in the design of scenarios to make positive and negative examples of each element clear, it is possible that participants may have considered all or part of each of the compound questions when responding. This issue is an important consideration, especially if applying the elements as criteria to classroom materials. For example, it is possible for text to be readable, but not comprehensible, particularly if a passage contains idioms or phrases that are culture-specific. To address this issue for measure revisions, these items could be simplified to improve clarity or could be presented as separate questions.

Another issue related to the measure design, discussed previously, is the misfit of the applied (scenario) items for which ‘not applicable’ was the correct response. This response option may have been used in instances where participants did not know or were uncertain of the answers, rather than to indicate that, within the scenario, an example (positive or negative) of the element was not given. Revising the scenarios to represent only positive and negative examples of each UDA element, or using a Likert-type scale for evaluation of accessibility based on each UDA element will be considered for measure revisions.

Limitations of the analyses. A few limitations are related to the analyses. First, although the data fit the 1PL IRT model adequately overall, there are limitations to using a one-parameter model. The 1PL IRT model measures only item difficulty, constraining slopes (item discrimination) and asymptotes (guessing). It is possible that item discriminations would vary if ‘freed up’, and that guessing may have occurred. Using a 2- or 3PL model would allow variability in item discrimination and/or guessing, respectively, and would likely yield significantly different trait levels.

Another limitation is the estimation procedure, joint maximum likelihood estimation (JMLE), used by Winsteps software for scaling items. Although there are advantages to this estimation procedure, including its applicability across IRT models and computational efficiency, notable disadvantages exist. These include biased and inconsistent parameter estimates, especially for fixed-length tests, which occurs because item and person parameters are estimated simultaneously. These estimates are not optimal for calculating standard errors and lead to difficulty in interpreting standard errors. In addition, JMLE does not provide estimates for perfect scores (items or persons)

and has little utility when comparing fit across models (Embretson & Reise, 2000). An alternative to analyzing data using Winsteps would be to use BILOG-MG, which uses marginal maximum likelihood estimation (MMLE) to estimate ability and item parameters. Embertson and Reise (2000) note that MMLE generally yields more consistent parameter estimates because the estimation procedure uses expected frequencies based on response pattern (rather than observed) and the estimation process is iterative (rather than simultaneous).

Directions for Future Research

Despite the limitations noted above, and given that, to date, examining and addressing teacher knowledge of UDA has yet to be explored, this study provided an initial step in the endeavor to extend the application of UDA principles to classroom assessments. By designing and attempting to validate the use of a measure of teacher knowledge of UDA, an effort was made to describe what teachers know about assessment accessibility issues through their application of seven UDA principles. The results of this study primarily provide information for further measure development and some limited initial evidence that supports the need for teacher professional development in this area.

This study sets the stage for additional research to explore (a) the design and delivery of a professional development curriculum for UDA, (b) additional uses of the measure (which would require additional validity evidence), including the use of the measure presented herein (or parts thereof) as a pre-/post-test to evaluate the effectiveness of professional development programs in terms of increased teacher knowledge and application of UDA, and (c) specific applications of UDA to classroom

assessments (including comparisons of student scores on UD and non-UD tests in various subject areas).

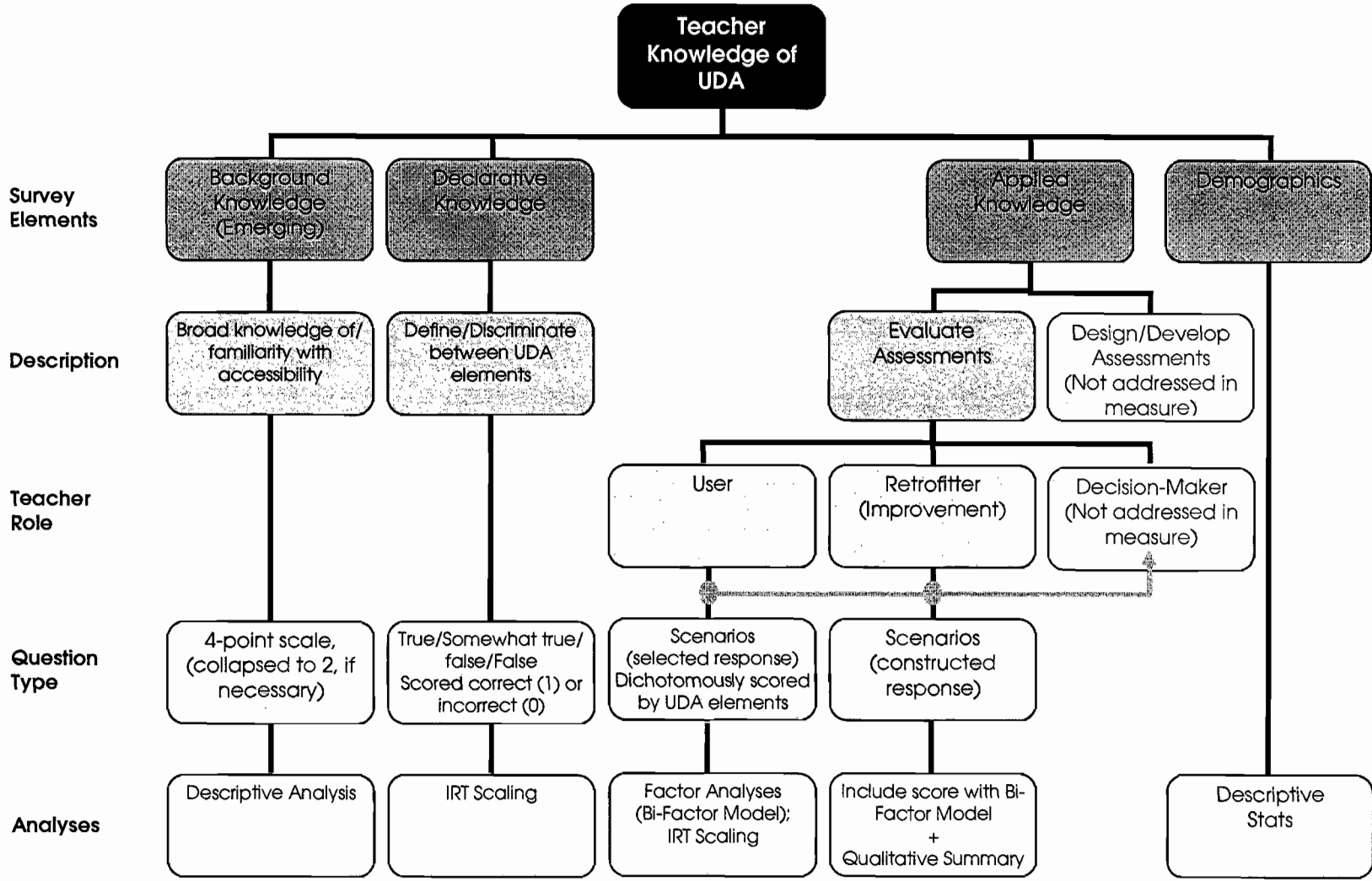
APPENDIX A
MEASURE BLUEPRINT

Measure Blueprint
Declarative and Applied Knowledge Items

UDA Principle	Declarative Knowledge (True/False) Items																				Applied Knowledge Items							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Tot.	1	2	3	4	5	6	7/8*
1. Inclusive assessment population	x	x				x															3	Y	N	Y	Y	N	N	
2. Precisely defined constructs							x		x				x								3	Y	Y	Y	NA	N	Y	
3. Accessible, non-biased items			x							x											2	Y	N	N	N	N	N	
4. Amenable to accommodations				x				x										x			3	NA	NA	Y	Y	NA	N	
5. Simple, clear & intuitive instructions and procedures													x						x	x	3	Y	Y	Y	NA	Y	N	
6. Maximum readability & comprehensibility					x						x					x					3	Y	N	Y	N	Y	Y	
7. Maximum legibility											x			x			x				3	Y	N	N	Y	N	Y	

Selected Response
 Y = Yes (accessible, positive example)
 N = No (inaccessible, negative example)
 NA = Not applicable to scenario (element is not described in the scenario)
***Constructed Response**
 Evaluated qualitatively

APPENDIX B
MEASURE OVERVIEW



APPENDIX C

RECRUITMENT EMAIL – INFORMED CONSENT

Dear [name],

Hi! My name is Elisa Jamgochian and I am a doctoral candidate in Educational Methodology, Policy, and Leadership at the University of Oregon. I am writing to invite you to participate in a study that will support my dissertation research.

I am interested in teacher knowledge of test accessibility issues. To research this, I have designed a survey that will help me better understand what teachers (from preservice to 'expert'; grades K-8) know about assessment accessibility issues through the application of seven universal design for assessment principles. The proposed use of the measure for this study is to describe teachers' knowledge of universal design for assessment and provide initial evidence for its usefulness in identifying training and professional development needs.

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed "to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

To participate, you are asked to complete a two-part survey. Part 1 includes items that address your experience working with students of various abilities and backgrounds, true/false statements about test accessibility, and basic demographic information. It is expected that Part 1 will take approximately 15 minutes to complete. Part 2 contains eight items for which you are asked to rate the accessibility of a test scenario given seven criteria (elements of Universal Design for Assessment). It is expected that Part 2 will take approximately 20-25 minutes to complete.

****IF YOU ARE INTERESTED IN PARTICIPATING, PLEASE EMAIL ME:**
ejamgoch@uoregon.edu. I will reply with a unique participant ID number and links to the survey.

Your participation is voluntary. Your decision to participate will not affect your relationship with the school or district. If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty.

If you don't wish to participate, simply do not complete the survey. Responses will be anonymous; data will be compiled using a random identification code. Completing and submitting the surveys indicates your agreement to participate.

The potential risks are minimal, as we make sure no one has access to your responses. Your random identification number will be linked to your email address on a secure server, in a location separate from survey response data. You will be compensated for your participation in the study and will be entered in a drawing. For your participation,

you will receive a \$10 gift card to Amazon. In addition, you will be entered in a drawing for one of four iPod shuffles. Your compensation will be sent within 2 weeks of the anticipated survey completion date. Drawing items will also be sent at that time.

Please keep this email in your files. If you have any questions about the study, please contact me at:

Elisa Jamgochian
Email: ejamgoch@uoregon.edu
Phone: 714-335-9195

Or, you may contact my advisor at:

Paul Yovanoff, Ph.D.
Phone: (541) 346-1495

If you have questions regarding your rights as a research subject, contact the Office for Protection of Human Subjects, University of Oregon, Eugene, OR 97403, (541) 346-2510. This Office oversees the review of the research to protect your rights and is not involved with this study.

Thank you for your interest and help with my dissertation study! I appreciate your participation and time.

Sincerely,
Elisa Jamgochian

APPENDIX D

RECRUITMENT EMAIL – FOLLOW-UP

Dear [name];

Thank you for your interest in participating in my dissertation study! Included in this email are your participant identification number and links to parts one and two of the test accessibility survey.

Your participant ID # is: XXXX

You will need to enter your participant ID # for each part of the survey.

Survey Links:

To link to Part 1, click here: [link to online survey]

To link to Part 2, click here: [link to online survey]

Please keep this email in your files. If you have any questions about the study, please contact me at:

Elisa Jamgochian

Email: ejamgoch@uoregon.edu

Phone: 714-335-9195

I appreciate your time and participation.

Sincerely,

Elisa

APPENDIX E

RECRUITMENT EMAIL – REMINDER

Dear [name];

A friendly reminder – if you are still interested and have not yet completed the test accessibility survey, please do so by [date].

I've included again in this email your participant identification number and links to parts one and two of the test accessibility survey so the information is readily accessible.

Your participant ID # is: XXXX

You will need to enter your participant ID # for each part of the survey.

Survey Links:

To link to Part 1, click here: [link to online survey]

To link to Part 2, click here: [link to online survey]

Please keep this email in your files. If you have any questions about the study, please contact me at:

Elisa Jamgochian

Email: ejamgoch@uoregon.edu

Phone: 714-335-9195

I appreciate your time and participation.

Sincerely,

Elisa

APPENDIX F
TK-UDA PART I

INTRODUCTION: (Including informed consent letter)

Thank you for your interest in participating in this study!

I am interested in teacher knowledge of test accessibility issues. To research this, I have designed a survey that will help me better understand what teachers know about assessment accessibility issues through the application of seven universal design for assessment (UDA) principles. The proposed use of the measure for this study is to describe teachers' knowledge of UDA and provide initial evidence for its usefulness in identifying training and professional development needs.

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed "to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

To participate, you are asked to complete a two-part survey. Part 1 includes items that address your experience working with students of various abilities and backgrounds, true/false statements about test accessibility, and basic demographic information. It is expected that Part 1 will take approximately 15 minutes to complete. Part 2 contains six items for which you are asked to rate the accessibility of a test scenario given seven criteria (elements of Universal Design for Assessment), and two items for which you are asked to provide suggestions to improve the given scenario. It is expected that Part 2 will take approximately 20-25 minutes to complete.

Your participation is voluntary. Your decision to participate will not affect your relationship with the school or district. If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty.

If you don't wish to participate, simply do not complete the survey. Responses will be anonymous; data will be compiled using a random identification code. Completing and submitting the surveys indicates your agreement to participate.

The potential risks are minimal, as we make sure no one has access to your responses. Your random identification number will be linked to your email address on a secure server, in a location separate from survey response data. You will be compensated for your participation in the study and will be entered in a drawing. For your participation, you will receive a \$10 Amazon gift card. In addition, you will be entered in a drawing for one of four iPod shuffles. Your compensation will be sent within 2 weeks of the anticipated survey completion date. Drawing items will also be sent at that time.

If you have any questions about the study, please contact me at:
Elisa Jamgochian
Email: ejamgoch@uoregon.edu
Phone: 714-335-9195

Or, you may contact my advisor at:
Paul Yovanoff, Ph.D.
Phone: (541) 346-1495

If you have questions regarding your rights as a research subject, contact the Office for Protection of Human Subjects, University of Oregon, Eugene, OR 97403, (541) 346-2510. This Office oversees the review of the research to protect your rights and is not involved with this study. Thank you for your interest and help with my dissertation study! I appreciate your participation and time.

Sincerely, Elisa Jamgochian

Footer (on each page of measure): Please click [here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Participant ID #

Please rate each of the following statements.

I am familiar with...

	Not at all	A little	Mostly	Very
the Americans with Disabilities Act (ADA)				
Section 504 of the Rehabilitation Act				
the ADA Standards for Accessible Design				
the concept of Universal Design (in general)				
the concept of Universal Design for Learning/Instruction				
the concept of Universal Design for Assessment				

Please rate each of the following statements.

Within the past 5 years, I have had experience teaching...

	None	Very little	Some	A lot
students with physical disabilities				
students with learning disabilities				
students with language disabilities				
English Language Learners (Students for whom English is not their native/primary language)				
students who are economically disadvantaged				

Please respond to the following statements.

	Not within the past 5 years	1-3 times per year	4-6 times per year	Monthly (or more frequently)
I participate in IEP meetings.				
I attend student support team meetings.				

I participate in training/professional development related to...

	Not within the past 5 years	1-3 times per year	4-6 times per year	Monthly (or more frequently)
students with physical disabilities				
students with learning disabilities				
students with language disabilities				
English Language Learners				
economically disadvantaged students				

What type(s) of training? (Check all that apply). [LOGIC CHAIN]

- School- or district-sponsored professional development/in-service
- University-sponsored professional development/in-service
- Publisher-sponsored professional development/in-service
- College/University course
- Online course (not university sponsored)
- Independent reading (books, articles, etc.)
- Other (Please list)

Please rate the following statements.

	Not at all	Somewhat	Mostly	Very
Our school is physically accessible to people with disabilities.				
My classroom is physically accessible to students with disabilities.				
The curriculum is accessible to all students.				

	Yes	No
Are you allowed to provide accommodations to students who do not have an IEP or 504 plan?		

Please rate the following statements.

In my teaching, I provide accommodations for...

	Never	Rarely	Sometimes	Frequently
class assignments				
class tests				
district tests				
state tests				

If appropriate, I allow any student...

	Never	Rarely	Sometimes	Frequently
extra time to complete assignments				
extra time to complete tests				
to complete tests in alternate settings				
to respond to assignments in a variety of ways/formats				
to respond to test questions in a variety of ways/formats				
to take alternate forms of tests				

	Never	A few times per year	Monthly	Weekly (or more frequently)
I use technology to support instruction				

I use technology in the following ways... (Check all that apply). [LOGIC CHAIN]

- Presenting lessons
- Grading/report cards
- Word processing students assignments
- Word processing students tests
- Creating/Maintaining class web site
- Browsing the internet for lesson plans
- Collaborating with others online (chat, message boards, etc.)
- Other (Please describe)

My students use technology for the following class-related purposes... (Check all that apply). [LOGIC CHAIN]

- Completing assignments
- Making presentations
- Taking tests
- Doing research (using CD-ROMs or software)
- Browsing the internet
- Collaborating with others online (chat, message boards, etc.)
- Other (Please describe)

The statements on each of the next four pages refer to designing and administering assessments. Please select very true, somewhat true, somewhat false, or very false for each statement.

Please rate each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
For accountability assessments (e.g., state or district tests), the population of students tested does not need to include every student				
Limiting the population of students to be tested is never appropriate.				
Accommodations (e.g., having test directions read aloud, writing directly in test booklet, testing in small group, breaks during testing, etc.) increase access to assessments				
One way to reduce bias in testing is to examine whether any test items are more difficult for students from different subgroups				

Please rate each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
There is no need to provide additional test accommodations for tests that are universally-designed				
Readability is often calculated by considering sentence length and number of difficult words, under the assumption that shorter sentences and easier words make text more readable				
Students with different abilities and skills should have the opportunity to demonstrate proficiency on the same content				
A well-designed assessment measures the intended target skills and concepts				

Please rate each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
'Construct' refers to what a student needs to be able to do in order to complete a test item				
The usefulness of test results is improved when test items are carefully developed and reviewed for bias				
Readability of a test is not affected by students' previous experiences, achievement, and interests				
Legibility refers to the capability of being deciphered with ease				

Please rate each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Clearly defined constructs (including the content, intent and purpose of the assessment) promote accurate decisions based on student performance				
Understanding test instructions and procedures is not dependent on a student's experience, knowledge, or current concentration level				
Illustrations do not complicate the use of assistive technology (including magnifiers, enlargement, etc.)				
It is possible to write a disorganized text, full of incomprehensible sentences and still obtain a good readability score				

Please rate each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Legibility applies only to text				
A goal of universal design is to facilitate the use of appropriate accommodations				
Providing simplified instructions to students who cannot understand how they need to respond invalidates a test				
When planning or evaluating test directions and procedures, it is important to consider whether or not students are able to work independently through a test				

About you...

Which best describes your current teaching role?

- Preservice/Student Teacher/Intern
- Teacher (grades K-8)
- University Instructor/Faculty
- Researcher (not university-affiliated)

Which of the following best describes your schools community?

- Rural
- Suburban
- Urban

What grade(s) do you currently teach? (Select all that apply).

- K
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Which subjects do you currently teach? (Select all that apply).

- Elementary (all/multiple subjects - continue to next question)
- Special Education
- Language Arts
- Mathematics
- History/Social Studies
- Science
- Health/Physical Education
- Arts (Visual, Performing)
- Foreign Language
- Other (Please list)

Which of the following best describe your credential? (Select all that apply).

- General Education
- Special Education
- Mild/Moderate Disabilities
- Moderate/Severe Disabilities
- Early Childhood
- ELL Endorsement (CLAD, BCLAD, etc.)
- Elementary/Multiple Subjects
- Secondary/Single Subject
- If single subject, please list subject endorsements

How many years have you taught (including this year)?

Please complete the following (up to highest degree earned).

Bachelors Major

Bachelors Minor

Masters Degree

Doctoral Degree

Are you... (please check one)

- Hispanic, regardless of race
- Black, not of Hispanic origin
- White, not of Hispanic origin
- Asian or Pacific Islander
- American Indian or Alaskan Native
- Biracial/multiracial
- Decline to state
- Other

Is English your first (native) language? (Please check one).

- Yes
- No
- Decline to state

Are you... (Please check one).

- Female
- Male
- Decline to state

APPENDIX G

PART I CONTACT INFORMATION FORM

Thank you for completing Part 1 of this survey! As a token of my gratitude for your time and effort, I would like to compensate you with a \$10 Amazon gift Card.

Please complete the following contact information to receive your gift card.

**Please note: This information is not connected in any way to your survey responses, and will be kept separate from survey response data on a secure server until the completion of this research (anticipated completion: June 2010).

Name

I prefer to receive my gift card via: [LOGIC CHAIN]

- Email
- Mail

Email Address

Mailing Address

APPENDIX H
TK-UDA PART II

INTRODUCTION: (Including informed consent letter)

Thank you for your interest in participating in this study!

I am interested in teacher knowledge of test accessibility issues. To research this, I have designed a survey that will help me better understand what teachers know about assessment accessibility issues through the application of seven universal design for assessment (UDA) principles. The proposed use of the measure for this study is to describe teachers' knowledge of UDA and provide initial evidence for its usefulness in identifying training and professional development needs.

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed "to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

To participate, you are asked to complete a two-part survey. Part 1 includes items that address your experience working with students of various abilities and backgrounds, true/false statements about test accessibility, and basic demographic information. It is expected that Part 1 will take approximately 15 minutes to complete. Part 2 contains six items for which you are asked to rate the accessibility of a test scenario given seven criteria (elements of Universal Design for Assessment), and two items for which you are asked to provide suggestions to improve the given scenario. It is expected that Part 2 will take approximately 20-25 minutes to complete.

Your participation is voluntary. Your decision to participate will not affect your relationship with the school or district. If you decide to participate, you are free to withdraw your consent and discontinue participation at any time without penalty.

If you don't wish to participate, simply do not complete the survey. Responses will be anonymous; data will be compiled using a random identification code. Completing and submitting the surveys indicates your agreement to participate.

The potential risks are minimal, as we make sure no one has access to your responses. Your random identification number will be linked to your email address on a secure server, in a location separate from survey response data. You will be compensated for your participation in the study and will be entered in a drawing. For your participation, you will receive a \$10 Amazon gift card. In addition, you will be entered in a drawing for one of four iPod shuffles. Your compensation will be sent within 2 weeks of the anticipated survey completion date. Drawing items will also be sent at that time.

If you have any questions about the study, please contact me at:
Elisa Jamgochian
Email: ejamgoch@uoregon.edu
Phone: 714-335-9195

Or, you may contact my advisor at:
Paul Yovanoff, Ph.D.
Phone: (541) 346-1495

If you have questions regarding your rights as a research subject, contact the Office for Protection of Human Subjects, University of Oregon, Eugene, OR 97403, (541) 346-2510. This Office oversees the review of the research to protect your rights and is not involved with this study. Thank you for your interest and help with my dissertation study! I appreciate your participation and time.

Sincerely, Elisa Jamgochian

Footer (on each page of measure): Please [click here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Participant ID #

Introduction:

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed “to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

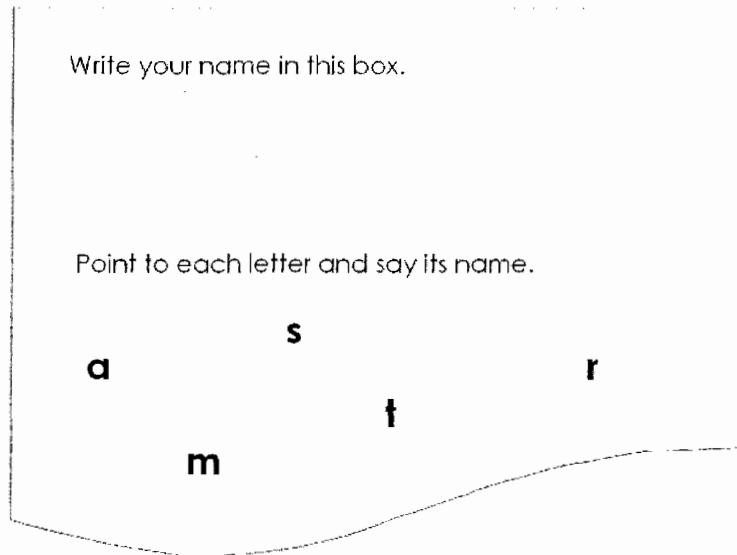
Directions:

For the following items, please indicate whether or not the test scenario and item presented are accessible, based on elements of Universal Design for Assessment given.

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario
 Click NO if a negative (poor) example of the element is given within the scenario
 Click N/A if the element is not present in the scenario (it is not applicable)

Setting: Each incoming kindergarten student at ABC School is screened prior to the start of the school year to assess his/her school readiness. A portion of the test is presented below. The teacher reads the test directions aloud and provides clarification to support student understanding.



	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario
 Click NO if a negative (poor) example of the element is given within the scenario
 Click N/A if the element is not present in the scenario (it is not applicable)

Setting: Each trimester, all second grade students at ABC School, except English Language Learners with less than one year in English Language instruction, are assessed to measure their progress in reading fluency. This is a timed, individually administered test. The teacher may read the directions aloud to each student.

Read the following excerpt aloud.

Now listen! In the country, close by the high road, stood a farmhouse: perhaps you have passed by and seen it yourself. There was a little flower garden with painted wooden palings in front of it; close by was a ditch, on its fresh green bank grew a little daisy; the sun shone as warmly and brightly upon it as on the magnificent garden flowers, and therefore it thrived well. One morning it had quite opened, and its little snow-white petals stood round the yellow centre, like the rays of the sun...

From H. C. Anderson's 'The Daisy'

	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario
 Click NO if a negative (poor) example of the element is given within the scenario
 Click N/A if the element is not present in the scenario (it is not applicable)

Setting: All fourth grade students participate in the district's Spring writing assessment. Students are asked to respond to a verbal/written prompt. Responses are scored for content, grammar, and organization. Students may use a dictionary or electronic translator and may write or type their responses.

Name _____

You will have 45 minutes to plan and write your response to the following prompt. You may use a dictionary or translator, and you may choose to write or type your response. Your score will be based on the content, grammar and organization of your response.

Writing Prompt: Write a description of an invention that would benefit humankind.

	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario

Click NO if a negative (poor) example of the element is given within the scenario

Click N/A if the element is not present in the scenario (it is not applicable)

Setting: Each sixth grade student is assessed in math prior to his/her placement in a middle school math course. Students may answer directly on the test document and use scratch paper, manipulatives and/or calculators to solve problems. A portion of the test is presented below.

Name _____

1. I am a number. To find out what I am you must take all of the digits in the largest four digit number and add them together. Divide that number by the minimum U.S. voting age and multiply the answer by itself. Take that number and divide it by the number of quarts in a gallon. Add the result to the number of items in a gross. Now you know what I am! What number am I?

	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario
 Click NO if a negative (poor) example of the element is given within the scenario
 Click N/A if the element is not present in the scenario (it is not applicable)


Setting: Middle school biology students participate in the districts general science assessment. The test is meant to assess students' knowledge of biology, chemistry, and physics, and contains multiple choice and short answer response formats. A portion of the test is presented below.

Name _____

PART 1: Circle the letter next to the correct answer.

Which of the following objects would have the most inertia?

a. Bowling ball
 b. Boulder
 c. Basketball
 d. Tennis ball



Which is an example of a mixture?

a. Garden
 b. Salad
 c. Salt
 d. Water

	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

DIRECTIONS: Given the information in the following scenario, please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment.

Click YES if a positive example of the element is given within the scenario
 Click NO if a negative (poor) example of the element is given within the scenario
 Click N/A if the element is not present in the scenario (it is not applicable)

Setting: Today, students in Mrs. Nelson's third grade reading group will be given a test to measure reading comprehension. Students are allowed to use a guide (e.g., a bookmark or blank sheet of paper) to follow along as they read. The teacher may not read any part of the test, including directions, to students.


Name _____

Directions: Circle the letter that corresponds to your answer. You might circle more than one letter to answer each question. Read each story carefully and answer the questions that follow.

1. Peter went to the baseball museum with his grandfather. They saw the Baseball Hall of Fame. Peter found a picture of his favorite player. For lunch, he ate food that is often sold at baseball games. His favorite was the apple pie!

What other food did Peter eat?

a. Hot dog
 b. Peanuts
 c. Burrito
 d. Ice cream



	Yes	No	N/A
Are the instructions and procedures simple, clear, & intuitive?			
Does the test/item demonstrate maximum readability & comprehensibility?			
Does the test/item demonstrate maximum legibility?			
Does the scenario represent inclusive assessment practices?			
Is the construct being measured by this test/item precisely defined?			
Are the test items accessible and non-biased?			
Is the test amenable to accommodations?			

For the next two scenarios, please describe how you would revise the test setting, directions, and items to improve accessibility.

DIRECTIONS: Please describe how you would revise this scenario (the setting, directions, and test item) to improve its accessibility.

Setting: Mr. Martin administers the district's 5th grade benchmark math assessment to his students each trimester. Students complete the test without help from their teacher or peers, and they are not allowed to use a calculator. They may write directly on the test, but need to transcribe their answers onto a scantron. A portion of the test is presented below.

Name _____

Directions: Fill in the bubble on your scantron that matches your answer. You may write on the test, but you may not use a calculator.

The Riddler has left a clue for Batman to follow at the scene of each crime. These are the clues that Batman has found:

- There is a 1 in the thousands place.
 - The digit in the tens place is 9 times the digit in the thousands place.
 - Multiply the digit in the thousands place by 2.
 - The digit in the ones place is a hand without a thumb.
 - The digit in the hundreds is 2 less than the number in the tens.
- Solve the riddle to find the number and help Batman stop the Riddler.

- a) 19224
- b) 29724

Setting

Directions

Test Item

DIRECTIONS: Please describe how you would revise this scenario (the setting, directions, and test items) to improve its accessibility.

Setting: ELL students in Mrs. Angeli's 7th grade class are given brief weekly reading comprehension assessments. Students read the passage and answer the questions that follow, using evidence to support their answers as needed. Students write directly on the paper. Students have 10 minutes to read each passage and respond to the questions.

Name _____

America's 78 million credit cardholders carried an average balance of \$7,564 last year. The cost in interest and fees amounted to more than \$1,000 for the typical budget. If you just said, "Budget - what budget?," you know what I mean. Truth is, most of us go on spending sprees from time to time. But, when power shopping creates the illusion of success, even as debts spiral out of control, it has become a weakness. Some obvious signs that spending is out of control include making minimum payments on your credit cards, late fees, bounced checks, lack of a budget and loss of sleep over money worries.

Answer the questions using information from the paragraph. Give evidence if needed.

- 1. How many people in America have credit cards? On the average how much do they put on their cards?*
- 2. What indicates you may be shopping more than you should?*
- 3. Do you feel as if the author may feel that shopping is bad for our health? Explain.*
- 4. Why did Terry's mom give Terry two ice cream sandwiches?*

Setting

Directions

Test Item

APPENDIX I

PART II CONTACT INFORMATION FORM

Thank you for completing Part 2 of this survey! Please complete the following contact information to be entered in a drawing to receive one of four iPod Shuffles.

**Please note: This information is not connected in any way to your survey responses, and will be kept separate from survey response data on a secure server until the completion of this research (anticipated completion: June 2010).

Name

Email Address

Mailing Address

APPENDIX J

TK-UDA PART I INTERNAL REVIEW FORM

INTRODUCTION

Thank you for reviewing this measure!

The purpose of the measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA) is to evaluate teachers' knowledge of test accessibility issues through their application of the seven elements of UDA. In general, the measure's content is based upon and derived from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999) to reflect a continuum/depth of knowledge.

The measure is comprised of two parts:

Part 1:

- Background Knowledge: Familiarity with federal acts and regulations related to accessibility, experiences working with students of various abilities and backgrounds, provisions for allowing student accommodations, and uses of technology.
- Declarative Knowledge: Statements reflect declarative (factual) knowledge of the elements of UDA. The content for these statements is based upon descriptions of each of the seven elements found in current research.
- Demographic Information: e.g., grades, subjects, and years taught, educational background

Part 2:

- Applied Knowledge: Six scenarios provide a description of a test setting and a sample student test item. For each scenario, participants evaluate the context (test setting and sample item) for accessibility using the seven UDA elements as their criteria. All student test items included in the scenarios are actual test items obtained from tests or student study materials available online. Participants are also presented with two additional scenarios for which they are asked to describe how they would revise the scenario to improve its accessibility in relation to test setting, directions, and sample item.

Part 1 of the TK-UDA is presented in its entirety on the following pages. On each page, there is a box for Reviewer Comments. Please provide any feedback that would help to improve the content and clarity of the items/measure.

Thanks again! I appreciate your time and support! -Elisa

Footer (on each page of review): Please click [here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Reviewer Initials

Please rate each of the following statements.

I am familiar with...

	Not at all	A little	Mostly	Very
the Americans with Disabilities Act (ADA)				
Section 504 of the Rehabilitation Act				
the ADA Standards for Accessible Design				
the concept of Universal Design (in general)				
the concept of Universal Design for Learning/Instruction				
the concept of Universal Design for Assessment				

Reviewer Comments

Please rate each of the following statements.

I have experience teaching...

	None	A little	Some	A lot
students with physical disabilities				
students with learning disabilities				
English Language Learners				
students who are economically disadvantaged				

Reviewer Comments

--

Please rate the following statements.

	Never	Rarely	Sometimes	Frequently
I participate in IEP meetings.				
I attend student support team meetings.				

I participate in training related to working with...

	Never	Rarely	Sometimes	Frequently
Students with physical disabilities				
Students with learning disabilities				
English Language Learners				
Economically disadvantaged students				

Reviewer Comments

--

What type(s) of training? (Check all that apply).

- Professional Development Workshop
- College/University course
- Online course (not university sponsored)
- Read books or articles
- Other (Please list)

*Note: this item is chained based on participants' responses to previous item.

Reviewer Comments

Please rate the following statements.

	Not at all	Somewhat	Fairly	Very
Our school is physically accessible to people with disabilities.				
My classroom is physically accessible to students with disabilities.				
The curriculum is accessible to all students.				

Reviewer Comments

Please rate the following statements.

I provide accommodations for...

	Never	Rarely	Sometimes	Frequently
Class assignments				
Class tests				
District tests				
State tests				

I allow any student...

	Never	Rarely	Sometimes	Frequently
extra time to complete assignments				
extra time to complete tests				
to complete tests in alternate settings				
to respond to assignments in a variety of ways/formats				
to respond to test questions in a variety of ways/formats				
to take alternate forms of tests				

Reviewer Comments

--

	Never	Rarely	Sometimes	Frequently
I use technology to support instruction				

Reviewer Comments

I use technology in the following ways... (Check all that apply).

- Presenting lessons
- Grading/report cards
- Word processing students assignments
- Word processing student tests
- Creating/Maintaining class web site
- Browsing internet for lesson plans
- Collaborating with others online (chat, message boards, etc.)
- My students use computers to complete assignments
- My students use computers to take tests
- My students use technology resources (Internet, CD-ROM, etc.) for research
- Other (Please describe)

*Note: this item is chained based on participants' responses to previous item.

Reviewer Comments

Please select true or false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
For accountability assessments (e.g., state or district tests), the target population does not need to include every student				
It is never appropriate to limit the population of students to be tested				
One way to reduce bias in testing is to examine whether any test items are more difficult for students from different subgroups				
There is no need to provide additional test accommodations for tests that are universally-designed				
Readability is often calculated by considering sentence length and number of difficult words, under the assumption that shorter sentences and easier words make text more readable				

Reviewer Comments

--

Please select true or false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Students with different abilities and skills should have the opportunity to demonstrate proficiency on the same content				
A well-designed assessment measures the intended target skills and concepts				
Accommodations increase access to assessments				
'Construct' refers to what a student needs to be able to do in order to complete a test item				
Careful item development and reviews of item bias improve the validity of test results				

Reviewer Comments

Please select true or false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Readability is not affected by students' previous experiences, achievement, and interests				
Legibility refers to the capability of being deciphered with ease				
Clearly defined constructs promote accurate decisions based on student performance				
Understanding test instructions and procedures is not dependent on a student's experience, knowledge, or current concentration level				
Illustrations do not complicate the use of assistive technology (including magnifiers, enlargement, etc.).				

Reviewer Comments

--

Please select true or false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
It is possible to write a disorganized text, full of incomprehensible sentences and still obtain a good readability score				
Legibility applies only to text				
A goal of universal design is to facilitate the use of the appropriate accommodations				
Simplified instructions invalidate a test taken by students who cannot understand how they need to respond				
An important consideration regarding test directions procedures is whether or not students are able to work independently through a test				

Reviewer Comments

--

About you...

Which best describes your current teaching role?

- Preservice/Student Teacher
- Teacher (grades K-8)
- University Instructor/Faculty

*Note: this item is chained based on participants' responses (e.g., Preservice teachers skip ahead to the item re: educational background).

Reviewer Comments

Which of the following best describes your schools' community?

- Rural
- Suburban
- Urban

Reviewer Comments

What grades do you currently teach? (Select all that apply).

- K
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Reviewer Comments

Which of the following best describe your credential? (Select all that apply).

- General Education
- Special Education
- Mild/Moderate Disabilities
- Moderate/Severe Disabilities
- Elementary/Multiple Subject
- Secondary/Single Subject
- If single subject, please list subject endorsements

Reviewer Comments

Which subjects do you currently teach? (Select all that apply).

- Elementary (all subjects - continue to next question)
- Language Arts
- Mathematics
- History/Social Studies
- Science
- Health/Physical Education
- Arts (Visual, Performing)
- Foreign Language
- Other (Please list)

Reviewer Comments

How many years have you taught (including this year)?

Reviewer Comments

Please complete the following (up to highest degree earned).

Bachelors Major

Bachelors Minor

Masters Major

Doctorate Major

Reviewer Comments

Are you... (please check one)

- Hispanic, regardless of race
- Black, not of Hispanic origin
- White, not of Hispanic origin
- Asian or Pacific Islander
- American Indian or Alaskan Native
- Biracial/multiracial
- Decline to state
- Other

Reviewer Comments

Is English your first (native) language? (Please check one).

- Yes
- No
- Decline to state

Reviewer Comments

Are you... (Please check one).

- Female
- Male
- Decline to state

Reviewer Comments

Any additional comments/suggestions?

APPENDIX K

TK-UDA PART I INTERNAL REVIEW COMMENTS

Section	Reviewer 1	Reviewer 2	Reviewer 3
1	these are perfect!		
2	excellent	do you want to quantify the 'teaching'. For example, 'in the last 5 years, I have experience teaching...'	
3	For the second question, I'm wondering if you want to place a time limit on when the training took place or takes place. For instance - I have participated in training in the past year related to working with...	On the "participate" do you want to specify a time range, such as "within the past year, I have participated...". Just a suggestion because someone could have participated 10 years ago and still call it participation. Or you could make your 1-4 ratings related to time such as 1-not within the past 5 years; 4-multiple times within a year For workshop, do you want to ask who sponsored it? For example, I remember attending "free" workshops offered by publishers that were just a sales pitch instead of university sponsored or district sponsored. Not sure if this would make things too complicated.	

Section	Reviewer 1	Reviewer 2	Reviewer 3
4	I like the additional information this will capture.		
5	Great questions.	Do you want to ask for more information? Perhaps this should come at the end of the survey?	Perhaps 'mostly' in place of 'fairly'
6	I like this accommodations section.	On the “any student” question, you may want to ask a question about the legal climate. Such as, are you able to provide accommodations to students who do not have an IEP. Someone may answer no to all of these questions but it isn't because they don't believe in it, but because they aren't allowed.	
7	Nice detail.	Do you want to specify types of technology?	'Word processing students assignments' I think should read: Word processing students' assignments same for students' tests Browsing [the] internet I would separate the 'I use technology in the following ways' and then add 'My students use technology in the following ways' setting apart the last 'my students' responses

Section	Reviewer 1	Reviewer 2	Reviewer 3
8		<p>The readability question seems a bit out of place. And it is a contentious issue. Do you need it?</p> <p>Ask some folks about the wording of these. I got stumped with “it is never appropriate...”. It just sounds funny to me.</p> <p>Also, the wording of #3 is a bit funny to me.</p>	<p>Directions should be consistent with response options (very true, somewhat true, ...</p>
9			<p>same as previous comment about directions</p>
10			<p>same</p>
11	<p>The true/false questions will provide you with excellent data.</p>		<p>'An important consideration regarding test directions procedures is whether or not students are able to work independently through a test'</p> <p>Phrase 'test directions procedures' is awkward. Consider rewording</p>
12			
13			<p>I must have missed that you are targeting K-8 dismiss previous comment</p>
14			<p>Multiple subject[s]</p>

Section	Reviewer 1	Reviewer 2	Reviewer 3
15			Perhaps this question could come before the last providing a definition of your use of the term 'subject'
16			as a certified teacher or taught at all, for example tutor? might want to clarify the use of the term 'taught'
17 – 20	[no comments]		
Additional Comments/ Suggestions	I like it. Well done!		

APPENDIX L

TK-UDA PART II INTERNAL REVIEW FORM

INTRODUCTION

Thank you for reviewing this measure!

The purpose of the measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA) is to evaluate teachers' knowledge of test accessibility issues through their application of the seven elements of UDA. In general, the measure's content is based upon and derived from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999) to reflect a continuum/depth of knowledge.

The measure is comprised of two parts:

Part 1:

- **Background Knowledge:** Familiarity with federal acts and regulations related to accessibility, experiences working with students of various abilities and backgrounds, provisions for allowing student accommodations, and uses of technology.
- **Declarative Knowledge:** Statements reflect declarative (factual) knowledge of the elements of UDA. The content for these statements is based upon descriptions of each of the seven elements found in current research.
- **Demographic Information:** e.g., grades, subjects, and years taught, educational background

Part 2:

- **Applied Knowledge:** Six scenarios provide a description of a test setting and a sample student test item. For each scenario, participants evaluate the context (test setting and sample item) for accessibility using the seven UDA elements as their criteria. All student test items included in the scenarios are actual test items obtained from tests or student study materials available online. Participants are also presented with two additional scenarios for which they are asked to describe how they would revise the scenario to improve its accessibility in relation to test setting, directions, and sample item.

Part 2 of the TK-UDA is presented in its entirety on the following pages. On each page, there is a box for Reviewer Comments. Please provide any feedback that would help to improve the content and clarity of the items/measure.

Thanks again! I appreciate your time and support! -Elisa

Footer (on each page of review): Please click [here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Reviewer Initials

Introduction:

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed “to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

Directions:

For the following items, please indicate whether or not the test scenario and item presented are accessible, based on elements of Universal Design for Assessment given.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each incoming kindergarten student at ABC School is screened prior to the start of the school year to assess his/her school readiness. The teacher reads the test directions aloud and provides clarification to support student understanding.

Write your name in this box.

Point to each letter and say its name.

a
s
r

m
t

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each trimester, all second grade students at ABC School, except English Language Learners with less than one year in English Language instruction, are assessed to measure their progress in reading fluency. This is a timed, individually administered test. The teacher may read the directions aloud to each student.

Read the following excerpt aloud.

Now listen! In the country, close by the high road, stood a farmhouse: perhaps you have passed by and seen it yourself. There was a little flower garden with painted wooden palings in front of it: close by was a ditch, on its fresh green bank grew a little daisy: the sun shone as warmly and brightly upon it as on the magnificent garden flowers, and therefore it thrived well. One morning it had quite opened, and its little snow-white petals stood round the yellow centre, like the rays of the sun...

From H. C. Anderson's 'The Daisy'

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: All fourth grade students participate in the district’s Spring writing assessment. Students are asked to respond to a verbal/written prompt. Responses are scored for content, grammar, and organization. Students may use a dictionary or translator and may write or type their responses.

Name _____

You will have 45 minutes to plan and write your response to the following prompt. You may use a dictionary or translator, and you may choose to write or type your response. Your score will be based on the content, grammar and organization of your response.

Writing Prompt: Write a description of an invention that would benefit humankind.

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each sixth grade student is assessed in math prior to his/her placement in a middle school math course. Students may answer directly on the test document and use scratch paper, manipulatives and/or calculators to solve problems.

Name _____

1. I am a number. To find out what I am you must take all of the digits in the largest four digit number and add them together. Divide that number by the minimum U.S. voting age and multiply the answer by itself. Take that number and divide it by the number of quarts in a gallon. Add the result to the number of items in a gross. Now you know what I am! What number am I?

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).


Setting: Middle School biology students participate in the state general science assessment. The test is meant to assess students' knowledge of biology, chemistry, and physics, and contains multiple choice and short answer response formats.

Name _____

PART 1: Circle the letter next to the correct answer.

Which of the following objects would have the most inertia?

a. Bowling ball
 b. Boulder
 c. Basketball
 d. Tennis ball



Which is an example of a mixture?

a. Garden
 b. Salad
 c. Soil
 d. Water

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Today, students in Mrs. Nelson's third grade reading group will be given a test to measure reading comprehension. Students are allowed to use a guide to follow along. The teacher may not read any part of the test, including directions, to students.


Name _____

Directions: Circle the letter that corresponds to your answer. You might circle more than one letter to answer each question. Read each story carefully and answer the questions that follow.

1. Peter went to the baseball museum with his grandfather. They saw the Baseball Hall of Fame. Peter found a picture of his favorite player. For lunch, he ate food that is often sold at baseball games. His favorite was the apple pie!

What other food did Peter eat?

a. Hot dog
b. Peanuts
c. Burrito
d. Ice cream



	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Comments

Please describe how you would revise this scenario (the setting, directions, and test item) to improve its accessibility.

Setting: Mr. Martin administers the district's 5th grade benchmark math assessment to his students each trimester. Students complete the test without help from their teacher or peers, and they are not allowed to use a calculator. They may write directly on the test, but need to transcribe their answers onto a scantron.

Name _____

Directions: Fill in the bubble on your scantron that matches your answer. You may write on the test, but you may not use a calculator.

The Riddler has left a clue for Batman to follow at the scene of each crime. These are the clues that Batman has found:

- There is a 1 in the thousands place.
- The digit in the tens place is 9 times the digit in the thousands place.
- Multiply the digit in the thousands place by 2.
- The digit in the ones place is a hand without a thumb.
- The digit in the hundreds is 2 less than the number in the tens.

Solve the riddle to find the number and help Batman stop the Riddler.

a) 19224
b) 29724

Setting

Directions

Test Item

Reviewer Comments

Please describe how you would revise this scenario (the setting, directions, and test items) to improve its accessibility.

Setting: ELL students in Mrs. Angeli’s 7th grade class are given brief weekly reading comprehension assessments. Students read the passage and answer the questions that follow, using evidence to support their answers as needed. Students write directly on the paper. Students have 10 minutes to read each passage and respond to the questions.

Name _____

America's 78 million credit cardholders carried an average balance of \$7,564 last year. The cost in interest and fees amounted to more than \$1,000 for the typical budget. If you just said, "Budget - what budget?" you know what I mean. Truth is, most of us go on spending sprees from time to time. But, when power shopping creates the illusion of success, even as debts spiral out of control, it has become a weakness. Some obvious signs that spending is out of control include making minimum payments on your credit cards, late fees, bounced checks, lack of a budget and loss of sleep over money worries.

Answer the questions using information from the paragraph. Give evidence if needed.

- 1. How many people in America have credit cards? On the average how much do they put on their cards? _____*
- 2. What indicates you may be shopping more than you should?*
- 3. Do you feel as if the author may feel that shopping is bad for our health? Explain.*
- 4. Why did Terry's mom give Terry two ice cream sandwiches?*

Setting

Directions

Test Item

Reviewer Comments

Any additional comments/suggestions?

--

APPENDIX M

TK-UDA PART II INTERNAL REVIEW COMMENTS

Reviewer 1	<p>These scenarios are very well thought out. You will capture a wide range of responses, depending upon the familiarity of UDA of the survey participant. This Part 2 is more challenging for the participant and will test their integrity related to responding honestly, rather than clicking to complete the survey. By capturing participants' written responses though in the scenarios related to changing the assessments you will force them into greater honesty. You should capture very good information here on participants' knowledge.</p> <p>The only confusion that may occur is that the scenarios start off with some non-examples of universal design and some less-familiar participants with UDA may not score the items with responses that reflect their true knowledge. Perhaps putting in the first scenario test items that use more standard font for example? But, you may have given more thought to this and have good reasons for the placement of the scenarios. Overall though, this looks real good.</p>
Reviewer 2	<p>Set off the directions with adding Directions in bold and perhaps underlining?</p>
Reviewer 3	[None]

APPENDIX N

TK-UDA PART I EXTERNAL/TEACHER REVIEW FORM

INTRODUCTION

Thank you for reviewing this measure!

The purpose of the measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA) is to evaluate teachers' knowledge of test accessibility issues through their application of the seven elements of UDA. In general, the measure's content is based upon and derived from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999) to reflect a continuum/depth of knowledge.

The measure is comprised of two parts:

Part 1:

- **Background Knowledge:** Familiarity with federal acts and regulations related to accessibility, experiences working with students of various abilities and backgrounds, provisions for allowing student accommodations, and uses of technology.
- **Declarative Knowledge:** Statements reflect declarative (factual) knowledge of the elements of UDA. The content for these statements is based upon descriptions of each of the seven elements found in current research.
- **Demographic Information:** e.g., grades, subjects, and years taught, educational background

Part 2:

- **Applied Knowledge:** Six scenarios provide a description of a test setting and a sample student test item. For each scenario, participants evaluate the context (test setting and sample item) for accessibility using the seven UDA elements as their criteria. All student test items included in the scenarios are actual test items obtained from tests or student study materials available online. Participants are also presented with two additional scenarios for which they are asked to describe how they would revise the scenario to improve its accessibility in relation to test setting, directions, and sample item.

Part 1 of the TK-UDA is presented in its entirety on the following pages. On each page, there is a box for Reviewer Comments. Please provide any feedback that would help to improve the content and clarity of the items/measure.

Thanks again! I appreciate your time and support! -Elisa

Footer (on each page of review): Please click [here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Reviewer Initials

Please rate each of the following statements.

I am familiar with...

	Not at all	A little	Mostly	Very
the Americans with Disabilities Act (ADA)				
Section 504 of the Rehabilitation Act				
the ADA Standards for Accessible Design				
the concept of Universal Design (in general)				
the concept of Universal Design for Learning/Instruction				
the concept of Universal Design for Assessment				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Please rate each of the following statements.

I have experience teaching...

	None	A little	Some	A lot
students with physical disabilities				
students with learning disabilities				
English Language Learners				
students who are economically disadvantaged				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

Please rate the following statements.

	Never	Rarely	Sometimes	Frequently
I participate in IEP meetings.				
I attend student support team meetings.				

I participate in training related to working with...

	Never	Rarely	Sometimes	Frequently
Students with physical disabilities				
Students with learning disabilities				
English Language Learners				
Economically disadvantaged students				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

What type(s) of training? (Check all that apply).

- Professional Development Workshop
- College/University course
- Online course (not university sponsored)
- Read books or articles
- Other (Please list)

*Note: this item is chained based on participants' responses to previous item.

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Please rate the following statements.

	Not at all	Somewhat	Fairly	Very
Our school is physically accessible to people with disabilities.				
My classroom is physically accessible to students with disabilities.				
The curriculum is accessible to all students.				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Please rate the following statements.

I provide accommodations for...	Never	Rarely	Sometimes	Frequently
Class assignments				
Class tests				
District tests				
State tests				

I allow any student...	Never	Rarely	Sometimes	Frequently
extra time to complete assignments				
extra time to complete tests				
to complete tests in alternate settings				
to respond to assignments in a variety of ways/formats				
to respond to test questions in a variety of ways/formats				
to take alternate forms of tests				

	Yes	No
Are you able to provide accommodations to students who do not have an IEP?		

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

	Never	Rarely	Sometimes	Frequently
I use technology to support instruction				

I use technology in the following ways... (Check all that apply).*

- Presenting lessons
- Grading/report cards
- Word processing students assignments
- Word processing students tests
- Creating/Maintaining class web site
- Browsing the internet for lesson plans
- Collaborating with others online (chat, message boards, etc.)
- Other (Please describe)

--

My students use technology in the following ways... (Check all that apply).*

- Completing assignments
- Presentations
- Taking tests
- Research (using CD-ROMs or software)
- Browsing the internet
- Collaborating with others online (chat, message boards, etc.)
- Other (Please describe)

--

*Note: this item is chained based on participants' responses to previous item (Sometimes or Frequently).

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Please select very true, somewhat true, somewhat false, or very false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
For accountability assessments (e.g., state or district tests), the target population does not need to include every student				
Limiting the population of students to be tested is never appropriate.				
One way to reduce bias is to examine whether any test items are more difficult for students from different subgroups				
There is no need to provide additional test accommodations for tests that are universally-designed				
Readability is often calculated by considering sentence length and number of difficult words, under the assumption that shorter sentences and easier words make text more readable				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

Please select very true, somewhat true, somewhat false, or very false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Students with different abilities and skills should have the opportunity to demonstrate proficiency on the same content				
A well-designed assessment measures the intended target skills and concepts				
Accommodations increase access to assessments				
'Construct' refers to what a student needs to be able to do in order to complete a test item				
Careful item development and reviews of item bias improve the validity of test results				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

Please select very true, somewhat true, somewhat false, or very false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
Readability is not affected by students' previous experiences, achievement, and interests				
Legibility refers to the capability of being deciphered with ease				
Clearly defined constructs promote accurate decisions based on student performance				
Understanding test instructions and procedures is not dependent on a student's experience, knowledge, or current concentration level				
Illustrations do not complicate the use of assistive technology (including magnifiers, enlargement, etc.).				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Please select very true, somewhat true, somewhat false, or very false for each of the following statements.

	Very true	Somewhat true	Somewhat false	Very false
It is possible to write a disorganized text, full of incomprehensible sentences and still obtain a good readability score				
Legibility applies only to text				
A goal of universal design is to facilitate the use of the appropriate accommodations				
Simplified instructions invalidate a test taken by students who cannot understand how they need to respond				
An important consideration regarding test directions procedures is whether or not students are able to work independently through a test				

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Does the scale represent an appropriate range of behaviors/responses?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

About you...

Which best describes your current teaching role?

- Preservice/Student Teacher
- Teacher (grades K-8)
- University Instructor/Faculty

*Note: this item is chained based on participants' responses (e.g., Preservice teachers skip ahead to the item re: educational background).

Reviewer Comments

Which of the following best describes your schools' community?

- Rural
- Suburban
- Urban

Reviewer Comments

What grades do you currently teach? (Select all that apply).

- K
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Reviewer Comments

Which subjects do you currently teach? (Select all that apply).

- Elementary (all subjects - continue to next question)
- Language Arts
- Mathematics
- History/Social Studies
- Science
- Health/Physical Education
- Arts (Visual, Performing)
- Foreign Language
- Other (Please list)

Reviewer Comments

Which of the following best describe your credential? (Select all that apply).

- General Education
- Special Education
- Mild/Moderate Disabilities
- Moderate/Severe Disabilities
- Elementary/Multiple Subject
- Secondary/Single Subject
- If single subject, please list subject endorsements

Reviewer Comments

How many years have you taught (including this year)?

Reviewer Comments

Please complete the following (up to highest degree earned).

Bachelors Major

Bachelors Minor

Masters Major

Doctorate Major

Reviewer Comments

Are you... (please check one)

- Hispanic, regardless of race
- Black, not of Hispanic origin
- White, not of Hispanic origin
- Asian or Pacific Islander
- American Indian or Alaskan Native
- Biracial/multiracial
- Decline to state
- Other

Reviewer Comments

Is English your first (native) language? (Please check one).

- Yes
- No
- Decline to state

Reviewer Comments

Are you... (Please check one).

- Female
- Male
- Decline to state

Reviewer Comments

Any additional comments/suggestions?

APPENDIX O

TK-UDA PART II EXTERNAL/TEACHER REVIEW FORM

INTRODUCTION

Thank you for reviewing this measure!

The purpose of the measure of Teacher Knowledge of Universal Design for Assessment (TK-UDA) is to evaluate teachers' knowledge of test accessibility issues through their application of the seven elements of UDA. In general, the measure's content is based upon and derived from federal acts and regulations (e.g., USDE: NCLB, 2001; IDEA, 2004), technical reports (e.g., Thompson, Johnstone, & Thurlow, 2002) and standards for fair, accurate and accessible tests (AERA, APA, NCME, 1999) to reflect a continuum/depth of knowledge.

The measure is comprised of two parts:

Part 1:

- **Background Knowledge:** Familiarity with federal acts and regulations related to accessibility, experiences working with students of various abilities and backgrounds, provisions for allowing student accommodations, and uses of technology.
- **Declarative Knowledge:** Statements reflect declarative (factual) knowledge of the elements of UDA. The content for these statements is based upon descriptions of each of the seven elements found in current research.
- **Demographic Information:** e.g., grades, subjects, and years taught, educational background

Part 2:

- **Applied Knowledge:** Six scenarios provide a description of a test setting and a sample student test item. For each scenario, participants evaluate the context (test setting and sample item) for accessibility using the seven UDA elements as their criteria. All student test items included in the scenarios are actual test items obtained from tests or student study materials available online. Participants are also presented with two additional scenarios for which they are asked to describe how they would revise the scenario to improve its accessibility in relation to test setting, directions, and sample item.

Part 2 of the TK-UDA is presented in its entirety on the following pages. On each page, there is a box for Reviewer Comments. Please provide any feedback that would help to improve the content and clarity of the items/measure.

Thanks again! I appreciate your time and support! -Elisa

Footer (on each page of review): Please click [here](#) to email Elisa Jamgochian if you have any questions regarding this survey.

Reviewer Initials

Introduction:

Universal design is a concept rooted in architecture and product design; at its core is the belief that products and environments can be designed “to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (Center for Universal Design, 2008, ¶ 1). Universal Design for Assessment (UDA) extends this concept to address issues of accessibility within assessment systems (Thompson, Johnstone, & Thurlow, 2002).

Directions:

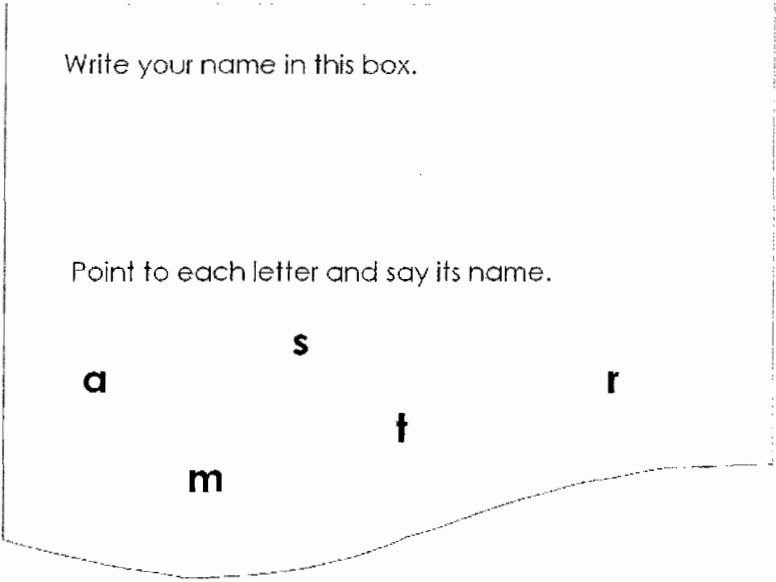
For the following items, please indicate whether or not the test scenario and item presented are accessible, based on elements of Universal Design for Assessment given.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each incoming kindergarten student at ABC School is screened prior to the start of the school year to assess his/her school readiness. The teacher reads the test directions aloud and provides clarification to support student understanding.



	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each trimester, all second grade students at ABC School, except English Language Learners with less than one year in English Language instruction, are assessed to measure their progress in reading fluency. This is a timed, individually administered test. The teacher may read the directions aloud to each student.

Read the following excerpt aloud.

Now listen! In the country, close by the high road, stood a farmhouse: perhaps you have passed by and seen it yourself. There was a little flower garden with painted wooden palings in front of it; close by was a ditch. on its fresh green bank grew a little daisy: the sun shone as warmly and brightly upon it as on the magnificent garden flowers, and therefore it thrived well. One morning it had quite opened, and its little snow-white petals stood round the yellow centre, like the rays of the sun...

From H. C. Anderson's 'The Daisy'

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: All fourth grade students participate in the district's Spring writing assessment. Students are asked to respond to a verbal/written prompt. Responses are scored for content, grammar, and organization. Students may use a dictionary or translator and may write or type their responses.

Name _____

You will have 45 minutes to plan and write your response to the following prompt. You may use a dictionary or translator, and you may choose to write or type your response. Your score will be based on the content, grammar and organization of your response.

Writing Prompt: Write a description of an invention that would benefit humankind.

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Each sixth grade student is assessed in math prior to his/her placement in a middle school math course. Students may answer directly on the test document and use scratch paper, manipulatives and/or calculators to solve problems.

Name _____

1. I am a number. To find out what I am you must take all of the digits in the largest four digit number and add them together. Divide that number by the minimum U.S. voting age and multiply the answer by itself. Take that number and divide it by the number of quarts in a gallon. Add the result to the number of items in a gross. Now you know what I am! What number am I?

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

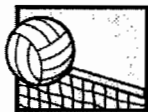
Setting: Middle School biology students participate in the state general science assessment. The test is meant to assess students' knowledge of biology, chemistry, and physics, and contains multiple choice and short answer response formats.

Name _____

PART 1: Circle the letter next to the correct answer.

Which of the following objects would have the most inertia?

a. Bowling ball
 b. Boulder
 c. Basketball
 d. Tennis ball



Which is an example of a mixture?

a. Garden
 b. Salad
 c. Spilt
 d. Water

	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

DIRECTIONS: Please indicate whether or not the test scenario and items presented are accessible, based on elements of Universal Design for Assessment listed below.

- If a positive example of the element is described within the scenario, click YES (it is accessible).
- If a negative example is given, click NO (it is not accessible).
- If the element is not described in the scenario, click N/A (the element is not applicable to the scenario).

Setting: Today, students in Mrs. Nelson’s third grade reading group will be given a test to measure reading comprehension. Students are allowed to use a guide to follow along. The teacher may not read any part of the test, including directions, to students.


Name _____

Directions: Circle the letter that corresponds to your answer. You might circle more than one letter to answer each question. Read each story carefully and answer the questions that follow.

1. Peter went to the baseball museum with his grandfather. They saw the Baseball Hall of Fame. Peter found a picture of his favorite player. For lunch, he ate food that is often sold at baseball games. His favorite was the apple pie!

What other food did Peter eat?

- a. Hot dog
- b. Peanuts
- c. Burrito
- d. Ice cream



	Yes	No	N/A
Simple, Clear, & Intuitive Instructions and Procedures			
Maximum Readability & Comprehensibility			
Maximum Legibility			
Inclusive Assessment Practices			
Precisely Defined Constructs			
Accessible, Non-Biased Items			
Amenable to Accommodations			

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Are the items clear and understandable?		
Are the response options appropriate?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

DIRECTIONS: Please describe how you would revise this scenario (the setting, directions, and test item) to improve its accessibility.

Setting: Mr. Martin administers the district's 5th grade benchmark math assessment to his students each trimester. Students complete the test without help from their teacher or peers, and they are not allowed to use a calculator. They may write directly on the test, but need to transcribe their answers onto a scantron.

Name _____

Directions: Fill in the bubble on your scantron that matches your answer. You may write on the test, but you may not use a calculator.

The Riddler has left a clue for Batman to follow at the scene of each crime. These are the clues that Batman has found:

- There is a 1 in the thousands place.
- The digit in the tens place is 9 times the digit in the thousands place.
- Multiply the digit in the thousands place by 2.
- The digit in the ones place is a hand without a thumb.
- The digit in the hundreds is 2 less than the number in the tens.

Solve the riddle to find the number and help Batman stop the Riddler.

a) 19224
b) 29724

Setting

Directions

Test Item

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

--

What, if any, misinterpretations might arise from the wording of the items/questions?

--

Any additional comments?

--

DIRECTIONS: Please describe how you would revise this scenario (the setting, directions, and test items) to improve its accessibility.

Setting: ELL students in Mrs. Angeli’s 7th grade class are given brief weekly reading comprehension assessments. Students read the passage and answer the questions that follow, using evidence to support their answers as needed. Students write directly on the paper. Students have 10 minutes to read each passage and respond to the questions.

Name _____

America's 78 million credit cardholders carried an average balance of \$7,564 last year. The cost in interest and fees amounted to more than \$1,000 for the typical budget. If you just said, "Budget - what budget?," you know what I mean. Truth is, most of us go on spending sprees from time to time. But, when power shopping creates the illusion of success, even as debts spiral out of control, it has become a weakness. Some obvious signs that spending is out of control include making minimum payments on your credit cards, late fees, bounced checks, lack of a budget and loss of sleep over money worries.

Answer the questions using information from the paragraph. Give evidence if needed.

- 1. How many people in America have credit cards? On the average how much do they put on their cards?* _____
- 2. What indicates you may be shopping more than you should?* _____
- 3. Do you feel as if the author may feel that shopping is bad for our health? Explain.* _____
- 4. Why did Terry's mom give Terry two ice cream sandwiches?* _____

Setting

Directions

Test Item

Reviewer Rating:

	Yes	No (Needs Improvement)
Are the directions clear and understandable?		
Does this scenario reflect a plausible classroom testing situation?		

If you selected No (Needs Improvement) for any of the above, please explain.

What, if any, misinterpretations might arise from the wording of the items/questions?

Any additional comments?

Any additional comments/suggestions (overall)?

APPENDIX P

TK-UDA EXTERNAL/TEACHER REVIEW COMMENTS

Section	Needs Improvement	Potential Misinterpretations	Additional
1	<p>I think it is good to have a 4-point scale. I am not sure about an equitable difference in the higher two 'somewhat' and Very. I believe 'a little' and 'somewhat' are close in similarity for descriptions.</p> <p>Just an idea/question. 'a little' and 'somewhat' are very similar. It seems like maybe you only need one of these options</p>	<p>Often UD is used for UDL/I--granted that isn't correct but I wonder if the UD question should be last, or at least after UDL/I so as to differentiate from UDL/I?</p> <p>Some people who know very little may choose somewhat because they don't want to look 'uninformed' aka stupid, so they may over rate their knowledge.</p>	<p>I was clicking on the last item's radio buttons just to see if it was forced response--I assume it is since I can't remove my button</p>
2	<p>I do prefer the language of this scale.</p> <p>it may be useful on the English Language Learners question to parenthetically write English is not native or primary language</p> <p>Could you change 'A little' to 'Very little' - that might eliminate some of the potential overlap between 'A little and some'</p> <p>Are you going to ask 'not sure' for any items? This may not be necessary, but is an idea.</p> <p>physical disability is differently interpreted in different states. maybe an i.e., would help?</p>	<p>lack of knowledge with current term of English Language Learners</p>	<p>The survey is a nice idea. My one concern is that whether or not teachers know a lot about UDA, they are stuck with the test they get (for large scale assessments). If this is for local or classroom-based assessments, this could be very informational!</p>

Section	Needs Improvement	Potential Misinterpretations	Additional
3	<p>I am struggling with the difference between 'rate' and perhaps 'respond'--aren't you really asking them to respond to the following here? You are asking for frequency, not rating of comfort or knowledge.</p> <p>Could you reword the opening statement??? It is wordy. ex. 'I have received training and/or continue to receive training related to'; 'I am trained to work with' or get rid of 'working with'</p>	<p>You could specialize each instruction page to improve clarity.</p>	<p>Excellent - very clear and understandable</p> <p>Is there a reason why only students with learning and physical disabilities were included?</p>
4	<p>Chained so if not in the last 5 years they don't get the item? If so then directions are fine.</p> <p>The item on Prof Dev may benefit from some clarification---under that would school, district and state workshops count? Or are you looking for something specific to the school/district. Also, if I am in an online or university course I HOPE that I read books/articles, so do you mean independent reading here? Or something like that?</p>		<p>My assumption is that the above question will be juxtaposed to the previous question. Even so, it may be useful to set a context of professional development training in education or for students with disabilities...</p>
5	<p>this one is tough, how can a physical entity be somewhat accessible for ALL users. I am implying all from the statement. in which case the 4 point scale is off. You are or are not accessible, somewhat or mostly is irrelevant. You can get in the door but not the classrooms...</p> <p>Could you say 'Not at all accessible', etc. instead of just 'Not at all'?</p>	<p>Yes--but I wonder if it might be helpful to have a sub-question set under the curriculum is accessible question--and have them respond specifically to the 4 items you asked earlier (PD, LD, ELL, E Disadvantaged)--wouldn't that give you more information? Or if you respond no to the question, a chained pop up as I described.</p>	

Section	Needs Improvement	Potential Misinterpretations	Additional
6	<p>Set the stage for the responses. I appreciate keeping the directions simple, here it may be useful to set up in my classroom, in my teaching, in our school...</p> <p>in the allow my student, set a context, is this in an assessment situation, general class day routine, assignments etc. or ?</p> <p>On the first section--all students or just those with a 504/iep.</p> <p>Final question--is this needed if they answer yes to the above section? And shouldn't 504 be also listed?</p> <p>Are you asking if the teacher is 'able' meaning capable or allowed?</p> <p>The 'I allow' question is a bit frustrating. I assume this is in general, but in some situations some people may not do these things. I.e. I can see people thinking that if the student doesn't use class time wisely, they don't provide extra time. Not sure what you want to know. Could you add 'when necessary'</p>	<p>It may be worthwhile to define what is meant by accommodation, unless you want to evaluate understanding</p>	<p>In the section 'I provide accommodations for...' you mention tests and assignments... 'projects' came to mind, but I guess that would be part of 'assignments'? Just a thought.</p>
7	<p>For the technology item, you may need to be more specific (e.g., never, once per month, weekly, daily)</p> <p>I am a big parallel structure person--so fix the 2nd chained set to reading making presentations, completing research....</p>	Nice	

Section	Needs Improvement	Potential Misinterpretations	Additional
8	<p>The directions are clear, but could possibly be simplified. Select the appropriate rating for the following statements.</p> <p>Not sure about the first one. Are the statements intended to catch teacher perceptions? If so, would a Likert scale around agree to disagree be better?</p> <p>Ok--if intentional leave it, but in the previous items you go from negative to positive and now from positive to negative---while it is good to have some 'truth' seeking items I hope that you don't get some incorrect responses here.</p> <p>This is a big shift from your previous instructions----don't you want them to rate here? Whatever you do, keep it parallel for the reader.</p> <p>What if someone doesn't understand what you mean by target population?</p>		<p>This was a really abrupt shift from the previous items--I know that you don't want to impact your responses by too much additional information, but a brief intro like The following items will ask you to respond to information about testing, or something like that to help the shift.</p>
9	<p>Same as previous on directions</p> <p>See previous page comments</p> <p>Same comment about the instructions.</p> <p>Should 'reviews' in the last item be 'review'</p> <p>I don't understand the last statement 'Careful item development...'</p> <p>By 'item' do you mean problem on a test??</p>		<p>I think somewhere you should ask if they know they definition for accommodations? And then maybe even define what you mean? When I work with working teachers, often modification/ accommodations are used interchangeably. They aren't...but this would skew your response data.</p>

Section	Needs Improvement	Potential Misinterpretations	Additional
10	<p>See previous directions comment. Otherwise very clear.</p> <p>See previous two page comments</p> <p>The first 2 items confused me---and this seems like a strange set of items.</p> <p>Student writing readability?</p> <p>Student writing legibility?</p> <p>Sorry I have trouble with 'clearly defined constructs...!' not sure what you're asking.</p> <p>I'm not sure that it is necessarily a problem with the item...it may just be that I don't know what it means...but I'm not sure what 'Clearly defined constructs promote accurate decisions based on student performance' means.</p>	<p>Some language issues depending on audience.</p> <p>Constructs is the biggy.</p>	
11	<p>See previous three page comments</p> <p>What constitutes text? A paragraph? A sentence?</p> <p>I have had to reread the following a few times and I am still not sure what you want: An important consideration regarding test directions and procedures is whether or not students are able to work independently through a test</p>		

Section	General Comments
12	<p>Not sure of participants, might you have researchers unaffiliated with a University. Like PIR or ORI or CAST...</p> <p>Will you survey at all three levels? If so, consider that access issues are different for universities than they are for K-12 education (e.g., students can't go to universities if they don't demonstrate a certain level of achievement, but all students can access K-12 education).</p>
13	<p>These may need to be defined. For example put population ranges behind Urban and suburban. Has a term been applied to smaller towns and cities that aren't a suburb of a lg pop'n center?</p> <p>I don't know if you possibly want population ranges in here. (greater than 50,000, etc) I can see some cities/areas being confused by this item.</p>
14	<p>This is fine, but didn't you have three levels (elementary, secondary, tertiary) [before]?</p> <p>What about combo? Do you want to know if it is a single class but combined grades?</p> <p>What about preK or sped?</p> <p>Straight forward.</p>
15	<p>Good list of choices</p> <p>You may wish to check on the credential categories in Oregon, they may be very different. And you should also check on the ages of an ECSE credential--for some it goes to g1 and may cover the early grades.</p> <p>What about admin, SPL, psych, counselor, etc?</p> <p>Do you want to include anything about the CLAD/BCLAD? or are those outdated?</p>
16	<p>Do you want to create a scale for ease of analysis? (e.g., 0-3 years, 4-7 years, etc.)</p>
17	<p>Do you want to create a scale for ease of analysis? (e.g., 0-3 years, 4-7 years, etc.)</p>
18	<p>Do you really have a major for a MS or PHD?</p>
19	<p>What does 'regardless of race' mean? Just curious</p>
20	<p>Well stated</p>
Other	<p>This is a very good survey. Well done. Thanks for letting me take a peek and comment.</p> <p>I hope my feedback is helpful!</p> <p>Very interesting survey! You have a lot of really good questions. I hope my input helps. Ignore what is not helpful! Good luck!</p> <p>Very clear for me. Nicely done.</p> <p>This looks great! I don't feel like I was all that helpful because I didn't have many comments, but this is because it is very clear and well done.</p>

REFERENCES

- Acrey, C., Johnstone, C., & Milligan, C. (2005). Using universal design to unlock the potential for academic achievement of at-risk learners. *TEACHING Exceptional Children, 38*, 22-31.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.
- Bremer, C. D., Clapper, A. T., Hitchcock, C., Hall, T., & Kachgal, M. (2002). Universal design: A strategy to support students' access to the general education curriculum. *Information Brief, 1* (3), 1-5.
- Center for Applied Special Technology (2008). *What is universal design for learning?* Retrieved November 5 from <http://www.cast.org/research/udl/index.html>
- Center for Applied Special Technology (2008). *UDL Editions*. Available from: <http://udleditions.cast.org/>
- Center for Universal Design. (2008). *Universal design principles*. Retrieved November 4, 2008 from http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html#top
- Council of Chief State School Officers. (2008). *School data direct: United States public schools and districts*. <http://www.schooldatairect.org/app/data/q/stid=1036196/llid=162/stllid=676/locid=1036195/catid=1013/secid=4600/compid=859/site=pes>
- Coltrane, B. (2002, November). English language learners and high-stakes tests: An overview of the issues. *ERIC Digest*.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.
- Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*, 81-112.

- Dolan, R. P., Hall, T. E., Banjerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *The Journal of Technology, Learning, and Assessment*, 3 (7). Available from <http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1058&context=jtla>
- Dolan, R. P., Rose, D. H., Burling, K., Harms, M., & Way, D. (2007). *The universal design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests*. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.
- Dolan, R. P. & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives*, 27, 22-25.
- Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Erlandson, R. (2002). Universal Design for Learning: Curriculum, Technology, and Accessibility. In P. Barker & S. Rebelsky (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2002* (pp. 484-490). Chesapeake, VA: AACE.
- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38, 915-945.
- Gibbons, R.D. & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423-436.
- Hanna, E. I. (2005). *Inclusive Design for Maximum Accessibility: A Practical Approach to Universal Design* (PEM Research Report 05-04). Iowa City, IA: Pearson Educational Measurement.
- Hitchcock, C., Meyer, A., Rose, D., & Jackson, R. (2002). Providing new access to the general curriculum: Universal design for learning. *TEACHING Exceptional Children*, 35, 8-17.
- Johnstone, C. J. (2003). *Improving the validity of large-scale tests: Universal design and student performance* (Tech. Rep. No. 37). Minneapolis, MN: National Center on Educational Outcomes.
- Kame'enui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: The prevention of academic learning problems*. Columbus, OH: Merrill.

- Kame'enui, E. and Simmons, D. (1999). *Toward Successful Inclusion of Students with Disabilities: The architecture of instruction*. (ERIC/OSEP Mini Library on Adapting Curricular Materials, Vol. 1). Reston, VA: ERIC Clearinghouse on Disabilities and Gifted Education.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Ketterlin-Geller, L. R. & Johnstone, C. (2006). Accommodations and universal design: Supporting access to assessments in higher education. *Journal of Postsecondary Education and Disability*, 19 (2), 163-172.
- Ketterlin-Geller, L. R. (Fall 2008). Testing students with special needs: A model for understanding the interactions between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, 27, 3-16.
- Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universal design for assessment. *Journal of Technology, Learning, and Assessment*, 4(2). Available from <http://www.jtla.org>
- Klingner, J. (2004). The science of professional development. *Journal of Learning Disabilities*, 37, 248-255.
- Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, (220), 583-621.
- Linacre, J. M. (2009). *A Users Guide to Winsteps Ministep Rasch-Model Computer Programs*. (Program Manual 3.69.0). Retrieved from: <http://www.winsteps.com/>
- Linacre, J. M. (2001). WINSTEPS Rasch measurement computer program (Version 3.31) [Computer software]. Chicago: Winsteps.com.
- Meo, G. (2008). Curriculum planning for all learners: Applying universal design for learning (UDL) to a high school reading comprehension program. *Preventing School Failure*, 52, 21-30.
- Menken, K. (2000). *What are the critical issues in wide-scale assessment of English language learners?* (Issue Brief No. 6). Washington, DC: National Clearinghouse for Bilingual Education.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence.

- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Orkwis, R. & McLane, K. (1998, Fall). A curriculum every student can use: Design principles for student access. *ERIC/OSEP Topical Brief*, 3-19.
- Paige, R. (2004). Secretary Paige issues new policy for calculating participation rates under No Child Left Behind. Washington, DC: [Author].
- Pisha, B. & Coyne, P. (2001). Smart from the start: The promise of universal design for learning. *Remedial and Special Education*, 22, 197-203.
- Rose, D. & Dolan, R. (2000). Universal design for learning. *Journal of Special Education Technology*, 15(4). Available from <http://jset.unlv.edu/15.4/asseds/rose.html>
- Rose, D. and Meyer, A. (2000). Universal design for learning, associate editor column. *Journal of Special Education Technology*, 15 (1). Available from: <http://jset.unlv.edu/15.1/asseds/rose.html>
- Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: ASCD. Retrieved October 10, 2007, from <http://www.cast.org/teachingeverystudent/ideas/tes/>
- Stevens, J. P. (2001). *Applied Multivariate Statistics for the Social Sciences* (4th Ed.). Mahwah, NJ: Erlbaum.
- Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18, 23-27.
- Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thompson, S. & Thurlow, M. (June 2002). Universally designed assessment: Better tests for everyone! *NCEO Policy Directions*, 14.
- Thompson, S.J., Johnstone, C.J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 22, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm>

- Thurlow, M. L., McGrew, K.S., Tindal, G., Thompson, S. L., Ysseldyke, J. E., & Elliott, J. L. (2000). *Assessment accommodations research: Considerations for design and analysis* (Technical Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved April 16, 2009, from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Technical26.htm>
- Tindal, G., & Fuchs, L. (1999). *A summary of research on testing accommodations: An empirical basis for defining accommodations*. Lexington, KY: University of Kentucky, Mid-South Regional Resource Center.
- U.S. Department of Education, Office of Elementary and Secondary Education. (2004). *No Child Left Behind*. Washington, DC: [AUTHOR].
- U.S. Department of Education, Office of Special Education and Rehabilitative Services. (2004). *Individuals with Disabilities Education Act*. Washington, DC: [AUTHOR].
- U.S. Department of Education. (2002). *President's Commission on Excellence in Special Education*. Available from: <http://www.ed.gov/inits/commissionsboards/whspecialeducation/reports/index.html>
- Wilson, D. T., Wood, R., & Gibbons, R. (2003). *TESTFACT: Test scoring, item statistics, and item factor analysis* (Version 4.0.2) [Computer Software]. Chicago: Scientific Software International.