UNIVERSITY OF CALIFORNIA
RIVERSIDE

Using Assessment to Inform Web Resource Selection: Development of the Usage Rating Profile-Web Resource (URP-WR)

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Arts

in

Education

by

Nina Rosalie Mandracchia

June 2020

Thesis Committee:
Dr. Wesley Sims, Chairperson
Dr. Kathleen King
Dr. Austin Johnson

The Thesis of Nina Rosalie Mandracchia is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Development and Preliminary Psychometric Evaluation of the Usage Rating Profile-Web Resource (URP-WR)

by

Nina Rosalie Mandracchia

Master of Arts, Graduate Program in Education
University of California, Riverside, June, 2020
Dr. Wesley Sims, Chairperson

With the ever-increasing use of technology, web-based resources are becoming more prominent. To support the identification, evaluation, and use of quality web-based resources, this study outlines the development of the Usage Rating Profile, Web-Resource (UPR-WR). This extends an existing library of URP tools. Initial items were generated to align with four usability considerations identified in available literature. Identified considerations included: *acceptability, appearance, credibility, and feasibility*. Content validation through a consensus building activity resulted in preliminary organization and item reduction. An exploratory factor analysis extracted four factors: *reasonability, acceptability, appearance, and systems support,* with reasonability containing items from hypothesized factors of *credibility* and *feasibility*. The extracted factors demonstrated acceptable reliability estimates (α = 0.82-0.93). The URP-WR also demonstrated an acceptable level of social validity as determined by subscales of the URP-A. Finally, limitations and future directions are discussed.

Table of Contents

List of Tables

List of Figures

Introduction

With the onset of shelter in place mandates surrounding the outbreak of COVID-19, technology use has skyrocketed. Since February 29, 2020, Facebook has seen a 27% increase in website traffic, with Netflix not far behind at a 16% increase (Koeze & Popper, 2020). Average screen time reports displayed to iPhone users have anecdotally increased as well. The average screen time used by adults was around 3.5 hours per day in 2019 (Andrews, 2020); Apple has not released the screen time increase since the start of shelter in place, but people are reporting anywhere between a 36%-185% increase in screen time per day as compared to before the shelter in place mandates (Andrews, 2020). Video conferencing platforms have also seen their usage explode; for example, Zoom went from an average of approximately 2 million sessions per day in February to over 7 million per day in March, with other online platforms seeing increases as well (Koeze & Popper, 2020). With technology dominating so much of 2020 American citizens lives, especially in crises such as the current pandemic, it is important to analyze the role that technology plays in more than just screen time.

In the last 20 years, use of web-based resources to identify, access, and utilize information has increased exponentially. As recently as 1993, over 99% of households had no internet access; however, only 23 years later (2016), the numbers almost flipped so that 88% of households had internet access (Fischer-Baum, 2017). Like society at large, educators, including school psychologists utilize web-based resources now more than ever (Cummings, 2011). Web-based technology, including search engines and sites, makes it possible to easily access and disseminate information. The sheer volume of

available information can overwhelm education professionals. Information overload, the phenomenon in which a consumer presented with too much information (i.e., more than six choices) demonstrates a struggle with or even an inability to make a decision or choice (Buchanan & Knock, 2001). In short, too much information too quickly, can be overwhelming, confusing, and problematic. For example, consider a timely topic in education, multi-tiered systems of support, or MTSS. A Google search for "MTSS" results in 2,900,000 hits in a span of 0.40 seconds (Google). Unfortunately, the volume of search may negatively impact user consumption and selection of available resources.

Analysis of internet usage, particularly relative to web resources, indicates that 53% of internet users select the top result (i.e. the first link; Miller, 2012). Intuitively, the first option should be the best option, but that is not always the case. Internet marketing strategies can manipulate search results so that the first option may be sponsored. To illustrate a hypothetical example, Foot Locker could pay Google a large sum of money to display Foot Locker's site as the first option when a key word phrase entered includes the word "foot." A consumer searching for a podiatrist would not benefit from this option. Additionally, companies can engage in what is known as Search Engine Optimization (SEO) in which they strategically align content with key word phrases that are most commonly searched (Pinkerton, 2000). There are many ways to accomplish this. For example, companies can hire contract workers or use AI technology to write hundreds of blogposts on their sites containing key phrases, they can also purchase domain names containing key word phrases that link back to the company's main site. Thus, when a consumer searches for that key phrase, the search engine's algorithm will likely direct the

consumer to the site of the company that invests time and effort into search engine optimization or sponsorship. While an important consideration when consuming and using information, ease of access should not be the only consideration as accessibility can be easily manipulated.

Internet Information Quality

Similarly, social media, internet marketing, and easily accessible web-page hosting/development services create the opportunity for dissemination of low-quality information. To this point, the medical field has thoroughly examined available medical resources. These evaluations found web-based medical resources often provided low quality information for topics such as breast cancer (Ream et al., 2009), HIV/AIDS (Benotsch et al., 2004), and cervical disk herniation (Morr et al., 2010). Though yet formally explored through empirical research, similar findings are likely to extend to the quality of disseminated educational information. The abundance of information, coupled with the potential for it to be of low-quality, puts a well-intentioned educator or school psychologist in a challenging position. Information overload and ease of dissemination may result in consumer resource use based on accessibility rather than quality.

School psychologists and education professionals are rarely trained on digital citizenship, a construct that ranges from protection of information (i.e., scam identification), detection of information quality, and safe and responsible use (e.g., cyberbullying, photo/video exchange, social media use; Ribble, 2012). While NASP emphasizes use of technology, the most recent NASP (2010) practice model's guidance does not include content related to evaluation of web-based resources or the concept of

information technology (e.g., selection of quality web-based resources for intervention implementation; NASP, 2010). This lack of emphasis in training, likely also found in the training of other education professions, leads to a gap in knowledge base concerning how to detect high quality information in a source that has not been peer reviewed. Johnson et al. (2019) found special educators reported a variety of issues with finding quality information including not knowing where to look, terminology changing, and lack of access to scholarly journals. Ideally, high-quality practices or procedures would include an empirical evidence-base while also being feasible. Unfortunately, the absence of one often renders a resource useless in spite of the other. An evidence-based practice or intervention (EBP/I) that is unattainable is of little value to users, while a feasible intervention that lacks an evidence base (i.e., evidence of efficacy) will likely fail to address the identified need (Noell & Gansle, 2014).

Importance of Evidence-Based Practice Use

Although a "definition" of evidence-based practice varies, four characteristics of EBP have been agreed upon by most organizations: "(a) the use of a sound experimental or evaluation design and appropriate analytical procedures, (b) empirical validation of effects, (c) clear implementation procedures, (d) replication of outcomes across implementation sites, and (e) evidence of sustainability" (Kerr & Nelson, 2006, p. 89). Interventions that meet these criteria are important because they have been shown to produce quantifiable results. Thus, the EBP/I movement instills a second method of quality control: the studies first must be peer-reviewed in order to be published, then they must be reviewed to ensure that they meaningfully contribute to the evidence base. The

necessity of EBP/I use in education is agreed upon almost universally in education (Kratochwill & Shernoff, 2003; Simonsen et al., 2008).

The push towards EBP/I use gained traction in the early 2000's, led by the EBI Task force and the work of Thomas Kratochwill (Kratochwill, 2002; Kratochwill & Shernoff, 2003). Historically, the primary source for EBP/I has been scholarly, empirical research. The primary means by which such information is available for on demand consumption is via peer-reviewed scholarly journals. While technology has attempted to make such outlets more easily accessible, scholarly journal articles remain difficult to access and less than consumer friendly for many practitioners (Cummings, 2011). Issues including fees associated with accessing journals, technical language used in writing, and frequent emphasis on methods and analyses rather than practical application of most research, perpetuate the research to practice gap (Johnson et al., 2019). Accessibility and consumption roadblocks for scholarly journals may result in school psychologists and education professionals using less than desirable web-based resources (Cummings, 2011).

Further evidence supporting the research to practice gap is illuminated by educational research in the areas of research utilization and implementation science (Huberman, 1994). Far too often, education professionals abandon the use of evidence-based practices claiming that they do not work or are not practical (Kratochwill & Shernoff, 2003). However, research demonstrates that the apparent failures of EBP/I may be more appropriately attributed to poor or incorrect implementation rather than the nature of the EBP/I itself (St. Peter Pipkin et al., 2010; Noell & Gansle, 2014). A number

of factors can negatively impact implementation fidelity including poor training, low treatment acceptability (i.e., social validity), and lack of support systems (e.g., training; Sterling-Turner & Watson, 2002; Noell & Gansle, 2014). One additional factor that may not be as readily apparent but may precede these factors relates to the identification and selection of supports and services to be implemented. Implementation fidelity may decrease when a practice that is identified and selected does not appropriately match with implementer knowledge and skills or to the resources available to support implementation. In some instances, these challenges could be related to system support, but they may also be related to poor or ill-informed user identification and selection. Given the challenges associated with accessing and consuming evidence supporting and outlining use of these practices, it is reasonable to believe these challenges would negatively impact implementation. Training, practice acceptability, and support systems are all directly related to the source of the information they are based on.

Improved EBP/I Dissemination

Some have attempted to bridge the gap between research and practice by collecting, summarizing, and organizing EBP/I to support their dissemination. Universities, State Department of Education Agencies, Federal Agencies, educators, scholars, professional organizations, and businesses are increasingly organizing information, resources, and products onto web-based platforms to support training and implementation (e.g., ies.gov/ncee/wwc, interventioncentral.org, nasponline.org, intensiveintervention.org, ebi.missouri.edu). While many such sites are appropriate sources of genuine EBP/I, others are not. Generally, such resources represent efforts to

support efficacious EBP/I use in schools, but these web-based resources do not have the benefit of evaluation to inform consumer use.

Research of Kratochwill (e.g., 2003) and others have shown the importance of using EBP/I, but the ease associated with adoption of the first web-based resource found on a site such as Pinterest can lead to users making ill-informed decisions. In a recent research utilization study, special education teachers indicated more frequent usage of resources from social media sources (e.g., Pinterest, Youtube, Facebook, Twitter) despite acknowledging these resources were likely less credible (Johnson et al., 2019). In contrast, the opposite pattern emerged for other quality-controlled sources (e.g., Intervention Central, National Center Websites, IRIS Center). Qualitative responses demonstrated an awareness of the higher credibility or trustworthiness for these resources, but less frequent consumption and lower levels of support to do so. This further illustrates the pressing need to bridge the gap between accessibility and quality of web-based resources. Frequently, the most effective problem-solving efforts are data driven. Data-based decision-making drives individuals to make more informed subjective decisions (Chafouleas, 2011). Application of data-based decision-making principles applied to web-based resources would seek to use data to make more informed subjective evaluations of these resources. Ultimately, these evolutions would lead to adoption of higher quality practices, those that are not only accessible, but are also evidence-based.

Importance of Assessment for Data-based Decisions

Use of data to inform decision making provides users with a quantifiable assessment of important characteristics of the object of examination (Chafouleas, 2011; Reynolds, 2010; Kamphaus & Frick, 2005). An underlying reason of assessment is to reduce subjectivity (Chafouleas, 2011; Hintze, Volpe, Shapiro, 2002). The assessment gives additional information (e.g., operational definitions, context) by which to evaluate the target subject/patient, object, construct, or behavior (Chafouleas, 2011; Hintze, Volpe, & Shapiro, 2002; Kamphaus & Frick, 2005). Furthermore, empirical assessment methodologies allow for the evaluation and accumulation of psychometric evidence (e.g., reliability and validity) to promote defensibility of results (Chafouleas, 2011) and normative comparisons. Similarly, aggregation of empirical evaluation by prior users could give potential users a mechanism to inform their decision making. For example, assessment could provide a user comparing two interventions additional information by which to make a selection decision. In this example, assessment information could give the user a quantifiable, valid, measure of the usability according to previous users. This use of aggregate information may be both easier to use and more relatable (i.e., socially valid) than information provided by quality-controlled sites like WWC. Although imperfect, the use of DBD in this manner is better than use of subjective evaluation that favors ease of access.

With these considerations in mind, an assessment designed to inform DBD around the identification and selection of web resources appears advantageous. Those charged with promoting EBP in schools, including school psychologists would greatly benefit

from an assessment tool designed to aid in evaluating web-based resources. Unfortunately, available assessments to support data-based decision making through evaluating the usability of web-based educational resources, appears limited.

Currently Available Evaluations of Web-based Resources

A review of publicly available assessments designed to evaluate web-based resources in general yielded few options (see Lydia M. Olson Library, 2018). The limited available tools for evaluating web-based resources appear to be designed specifically for educators (Schrock, 2019) are often found on university websites. For example, Northern Michigan University's Lydia M. Olsen Library (2018) provides an "Evaluating Internet Sources" page. This page provides students questions to use when evaluating a web resource across six criteria: authority, accuracy, objectivity, currency, coverage, and appearance. For example, questions ask: "Is it clear who is responsible for the contents of the page?" or "Does the content appear to contain any evidence of bias?"

Similarly, Shrock (2019) provides a number of checklists for evaluating educationally inclined web-based resources. An "ABCs" of website evaluation is provided, with a key component identified for each letter of the alphabet. Some examples of these include authority, efficiency, and verifiability. The tools provided on this website vary in format, but all contain questions that provide dichotomous answers (yes/no). Example questions include: "Does the page take a long time to load?" and "Does the information appear biased?" After answering the dichotomous questions, evaluators are asked to provide a narrative summary of their evaluation to compare sites, though specific guidance for making comparisons is not provided.

While well-intentioned, several potential usability challenges are evident for these evaluations. Currently available evaluations appear to be formatted in a manner consistent with a guide or checklist rather than a quantitative assessment. Users are tasked with answering a series of dichotomous questions (i.e., Yes or No) and left to interpret the significance of the presence of responses independently based on a narrative summary. Although dichotomous responses can be helpful to ask oneself when evaluating a resource, they do not provide the precision associated with a Likert scale in a quantifiable measure of the resource (Greenwald & O'Connell, 1970). This precision can guide decision making for a user looking for certain qualities in a web-based resource that would not be fully captured by dichotomous responses. Further, Likert scales have been demonstrated to gain the same information in fewer questions; thus, an evaluation tool using a Likert scale is more user friendly and potentially socially valid (Greenwald & O'Connell, 1970). In short, these checklists appear to lack formatting and psychometric evidence to support reliable and valid comparisons or diagnostic decisions. Finally, concerns are noted in the length of these evaluations. For example, in some instance, these evaluations include a web page for each area of consideration. Such formatting appears potentially complicated, difficulty to aggregate, and time consuming.

Additionally, other concerns associated with these tools relate to background knowledge needed for completion. A consistent theme across available tools focuses on asking raters to determine whether or not resource developers are credible (i.e., question related to the authority of the authors). If users are seeking out resources on a topic to better understand the topic, it stands to reason they are likely unaware of who is and is

not an expert on said topic. The apparent subjectivity of such evaluation components raises further concerns for the validity of these instruments.

While these assessments represent good faith attempts to evaluate web-based resource, several shortcomings are noted for these tools. Noted concerns and absence of reliability and validity evidence suggest their use to drive data-based decisions may be limited. In order to address these concerns, development of a psychometrically valid and reliable tool to evaluate online resources is needed.

Usability Assessment

In contrast to the dichotomous survey format describe previously, some researchers have applied psychometric analysis in the assessment of usability. The Usage Rating Profile (URP) assessment methodology applies a Likert rating scale assessment approach to objectively evaluate perceptions of usability. A wealth of research now documents the successful application of URP assessment methodology to assessment and intervention applications in variety of topics in education (see Chafouleas et al., 2009; Chafouleas et al., 2012). Acceptability of interventions and assessments is typically used as the primary facet to guide decision making. URP tools provide a quantifiable measure to drive selection that diversifies the criteria beyond acceptability, often the most heavily weighted factor in selection (Chafouleas et al., 2009).

For example, the User Rating Profile-Intervention (URP-I) was developed to access usability of educational interventions across domains including: *acceptability, understanding, feasibility, and systems support*. Items based on the hypothesized factors were generated and judged by an expert panel who rated the category the item belonged

to, the confidence in their rating, and the relevance of the item to the construct. Decision rules were made, and the items were cut from 78 to 55. The remaining 55 items were placed on a 6-point Likert scale. The underlying factor structure of these items was analyzed using 254 surveys filled out by undergraduate students using a provided vignette of an intervention. Further items were deleted based on decision rules. The factor structure revealed four factors, acceptability, knowledge, feasibility, and systems support. The factors showed strong reliability estimates (ranging from 0.84-0.96) and correlations amongst each other with the exception of systems support.

The other available tools on the URP website include the URP Supporting Students' Behavioral Needs (URP-NEEDS), the URP-Assessment (URP-A), the URP-Intervention, Revised (URP-IR) and Children's Usage Rating Profile (CURP; UCONN, 2020). The available URP forms range from 21-29 items, and are accompanied by scoring guides that outline which items correspond to the factors measured. There are no cut scores provided, or guidelines for what constitutes a good, medium, or bad score. Unfortunately, current URP forms do not provide a rating profile score for recommendations from web-based resources, which may be a promising novel application of URP assessment formatting and methodology.

IUA Approach to Validity

Development of new assessment may be better framed as a development *and* validation process, as development often occurs while anticipating a series of steps to accumulate validity evidence. Kane's (2013) addition of key phases/inferences of establishing validity allows for prioritization of types of evidence based on assessment

type. The first step in the interpretation/use argument (IUA) outlined by Kane (2013) is to clearly state the proposed interpretation or use of the assessment, (Kane, 2013, p. 2). The purpose of the IUA is to "make the reasoning inherent in proposed interpretations and uses explicit so that it can be better understood and evaluated." In other words, to identify the "what" the assessment is designed to reveal about the examinee (or object of examination) and what can be validly interpreted from this revelation (Kane, 2013, p. 10).

Step 2 evaluates the inferences made in the IUA by testing these claims empirically, starting with the most questionable assumptions (Cook et al., 2015). The four types of inferences made in an assessment as defined by Kane are: (a) scoring (assigning an accurate, reproducible quantifiable score to an observation), (b) generalization (obtaining a representative sample so that the assessment can be applied across multiple possible scenarios accurately), (c) extrapolation (interpreting the scores from the representative sample as applying to real life scenarios validly) and (d) implication (making a decision based on the interpretation from the extrapolation step; Cook et al., 2015). Each of these inferences have multiple methods of empirical testing, and validation of an assessment requires multiple studies in order to be complete.

The empirical study to be performed first is determined according to the weakest inference as well as Cronbach's (1989) criteria: (a) prior uncertainty, (b) information yield (the amount of knowledge the researchers stand to gain as a result of this empirical analysis), (c) cost, and (d) leverage (the importance of answering this question in order to convince test developers that the assessment is valid) (p. 165). The weakest inference

generally starts with scoring in a new assessment, as it is difficult to evaluate the generalizability, interpretation, or decision making of an assessment that does not yet have a quantitative score (Cook et al., 2015). This is followed by generalization or extrapolation and concluded with implication. After those have been tested and validated, the process moves to the next most questionable inferences until all inferences have been addressed and appropriately validated.

This Study

To address the limited availability of measures used to evaluate the quality of web-based resources, this study begins the development, refinement, and validation process for the Usage Rating Profile-Web Resource (URP-WR). URP-WR development sought to objectively evaluate web-based resources across four hypothesized domains: accessibility, appearance, credibility, and feasibility. The URP-WR extends the Usage Rating Profile (URP) formatting and assessment methodology first established by Witt & Martens (1983) to web-based resources. URP assessments are publicly available, and variations are now used to evaluate and inform user perceptions and adoption of educational intervention and assessment practices (see Chafouleas et al., 2009; Chafouleas et al., 2012).

URP literature outlines the process for validating a perception-based evaluation tool. A pool of initial items designed to measure internal and external factors relating to the source being evaluated is created, then narrowed down using content validation procedures and exploratory factor analyses, depending on the strength of the underlying theory. Although no IUA is explicitly provided for any of the URPS, they do state

purposes for example, “the purpose of the Children’s Usage Rating Profile (CURP) was therefore developed to assess those factors likely to influence intervention usage from the perspective of a student” (Briesh & Chafouleas, 2009). In other words, the scores can be interpreted in that a high score means the intervention is more likely to be used and a lower score meaning it is less likely to be used. The URP tools were found to have varying numbers of underlying factors (ranging from 3-6) being measured, with acceptable factor loadings and statistical significance. Thus, the URP tools have begun the process of psychometric validation ensuring that the first inference, scoring is met.

Like URPs before it, development of the URP-WR seeks to inform and guide consumer use of web-based resources, the implications of which are far reaching and long lasting. The URP-WR will allow school psychologists and education professionals a means of evaluating the usability of web-based resources used in the adoption and training of school-based student support practices. Furthermore, such an assessment could support decreasing the persistent research to practice gap in education that has loomed for decades (Carnine, 1997; Greenwood & Abbot, 2001; Cummings, 2011). The URP-WR could also serve to inform web-resource developers as they create and disseminate these resources.

The goals of this study are twofold. This study presents (a) initial development processes and (b) refinement processes for the URP-WR. These initial content validity and reliability evaluations address specifically the assumptions underlying the scoring inferences related to the assessment validation as outlined by Kane (i.e., IUA; 2013). Specifically, this study presents initial content validity evidence and reliability evidence.

Based on review of existing literature on the various influences on web-resource usage, items were generated using a hypothesized four construct structure (i.e., accessibility, appearance, credibility, and feasibility.) Next, this study provided an initial evaluation of the factor structure of the URP-WR. Exploratory Factor Analysis (EFA) guided URP-WR refinement through item inclusion and exclusion and grouping by identified factors. Additionally, acceptable levels of inter-rater agreement were anticipated during initial use of the URP-WR. Finally, the social validity of the URP-WR is explored. Acceptable levels of social validity are expected in regard to the URP-WR. Specific research questions include: 1) Will a hypothesized factor structure hold (i.e., research identified 4 likely factors)? 2) What item combinations are most appropriate to maximize URP-WR utility (i.e., valid results using fewest number of items possible)? 3) Does the URP-WR demonstrate reliability through internal consistency of items within factors? 4) Do users perceive this measure as usable?

Method

Participants

This study occurred across two broad stages: (a) initial item development and content validation and (b) pilot administration.

Development and Initial Content Validation

Participants were recruited from the University of California, Riverside (UCR) faculty and students in the School Psychology program. Eight participants completed the consensus building task. See Appendix A for the consensus building task. These included six students and two faculty in the School Psychology program.

Factor Analysis and Item (Data) Reduction

Participants included 94 faculty, in-service educators, and undergraduate and graduate students in the fields of education and psychology. Participants were recruited through emails and announcements in classes, research lab meetings, and social media. The majority of participants were female (n = 76) and Hispanic (n = 38). The majority of participants were students (n = 62) studying Education (n = 50). There were also a fair number of teachers (n = 20). The average age of participants was 29, but the majority of participants fell in the 18-22 age range (n = 42). See Table 2.1 for a breakdown of demographic information.

This study sought five participants per item, the ratio precedented in Chafouleas et al., (2009). However, sample size recommendations for EFA vary in available literature across several variable considerations. As factor analytic techniques began to emerge, some authors recommend sample sizes ranging from a 3:1 ratio to a 10:1 ratio of participants to items, others recommended a minimum of 100 participants to conduct factor analysis (Cattell, 1978; Nunnally, 1967; Kline, 1994). More recently, a sample size "rule of thumb" has been elusive. De Winter et al., (2009) showed that the minimum of 50 participants needed to interpret an exploratory factor analysis may be arbitrary as reliable results were found with less than 50 participants in simulated analyses with high levels of communality. An item to factor ratio of at least 7 has been suggested as a method of alleviating communality concerns as this ratio produces valid results with sample sizes over 150 (Mundfrom et al., 2005). Thus, the current sample size is small by most criteria. The study aims to collect more participants to alleviate this.

Social Validity Measure

A total of 75 participants elected to participate in the social validity part of the study. The majority of participants were female (*n* = 62) and Hispanic (*n* = 30). The majority of participants were students (*n* = 58) studying Education (*n* = 38). There were also a fair number of teachers (*n* = 19). The average age of participants was 28.66, but the majority of participants fell in the 18-22 age range *(n* = 30). See Table 2.1 for a breakdown of demographic information.

Table 2.1

Demographic Characteristics of Participants

Category	Subcategory	*n*	*n*
		EFA	URP-A
Gender	Female	76	62
	Male	18	13
Ethnicity	Hispanic	38	30
	White (not Hispanic)	25	22
	Biracial/Multiracial	9	7
	Asian or Pacific Islander	15	12
	Middle Eastern	4	4
	Other	2	1
Age	18-22	42	30
	23-29	28	24
	30+	21	21
Area of Study (students)	Education	50*	38*
	Psychology	12	5
	Other	15	10
Job function (professionals)	Teacher	20*	19*
	Aide/Assistant	5	3
	Professor	3	1
	Other	28	20

*Some participants identified as both students (marking an area of study) and teachers

Measures

URP-WR

The initial URP-WR included 70 items organized across four hypothesized domains encompassing usability. See Appendix B for the URP-WR pilot form. Items were formatted as statements capturing aspects of usability across appearance, accessibility, credibility, or feasibility. For example, "The resource cites its original sources." Raters respond to statements using a 6-point Likert scale anchored by a scale using strongly disagree to strongly agree (i.e., 1 = strongly disagree). This formatting is consistent with that of the other available URP tools.

URP-A

The URP-A is a 28-question evaluation of social validity of an assessment. See Appendix C for the URP-A. Participants will be asked to fill out the URP-A after completion of the URP-WR to measure social validity or acceptability of the URP-WR as a tool. The URP-A has a six-factor model, with acceptable model fit thus demonstrating validity in scoring. The alpha coefficients ranged from .71-.90, having acceptable reliability with the exception of one factor (system support $\alpha = .63$; Miller et al., 2013).

Procedures

The development and initial validation followed a four-step process. First, IUA development of the URP-WR was initiated, with scoring being the inference tested in this study. Preliminary item development, then preliminary content validation in the form of the consensus-building task were completed. Finally, the initially developed items were

administered with a provided web-based resource in order to collect data for an exploratory factor analysis.

URP-WR IUA

In an arguments-based approach to validation (Kane, 2013), assessment development begins by clearly outlining an argument around its intended interpretations and uses (Kane, 2013). Given the noted shortage in assessments of web-based resource usability, the URP-WR was developed to support consumer identification and selection of educational practices and interventions. Specifically, it will provide the consumer a metric by which to compare web-resources in a more empirically based manner than currently available tools provide. This should in turn increase the use of EBP/I by including factors such as credibility and feasibility (discussed in detail in the next section) as opposed to focusing on accessibility and appearance. If the logic continues to follow, this should lead to an increase in student outcomes due to the selection of EBP/I which are implemented with fidelity due to quality of the web-resource provided recommendations. Finally, the existence of the URP-WR will provide web-resource developers a quantifiable guide to use during development of a resource concerning an EBP/I. This will lead to the presence of more high-quality resources to be selected by the consumer.

The IUA for the URP-WR is to effectively evaluate web-based resources promoting EBP/I for usability of the resource itself as well as implementation feasibility of the proposed recommendations based on user's perceptions. In other words, if the user perceives the resource to have high levels of accessibility, pleasing appearance,

credibility, and feasibility, the URP-WR score will be high. This is important because it gives a quantifiable score associated with a user's perception, that can be used to make an informed decision about selection of resources.

Development

The development of the URP-WR began with a literature review to determine the factors underlying a quality web-based resource. The scarcity of peer-reviewed research on characteristics underlying usable web-based resources, necessitated inclusion of non-peer-reviewed, publicly available information in this review. Characteristics were drawn largely from a variety of resources including those provided by Kathy Schrock, university library websites, and existing URP assessment tools. The limited scholarly attention and related guidance on this topic further illustrated the necessity of research in this area. Similarly, the information that is available has not been empirically validated, and neglects research utilization as well as implementation fidelity concerns which are crucial to EBP/I.

To address these concerns, a literature review was conducted to determine essential factors of web-based resource evaluation. Ultimately, information gleaned from various sources identified several common characteristics associated with WR usability. Common characteristics identified included authorship, credibility, reliability, appearance (e.g., aesthetically pleasing), organization, accessibility, feasibility, technical components, and more. Identified trait or characteristic considerations for usable web-based resources were then grouped by commonality into broad themes. Noted

commonalities resulted in identification of four usability domains: appearance, accessibility, credibility and feasibility.

Appearance. Characteristics consistent with appearance included visual appeal, organization, use of pictures, use of headings, use of advertisements, size of font, and more. These characteristics were combined to encompass "appearance" which includes the aesthetic appeal as well logical organization of the resource.

Accessibility. Characteristics consistent with accessibility included the ease of finding the resource, ease of using the resource, length of time needed for the resource to load, presence of different modalities (e.g. option to read or listen to the information presented), presence of cost associated with accessing the resource, and more. These characteristics were combined to encompass "accessibility" which includes the ease associated with accessing and utilizing the resource.

Credibility. Characteristics consistent with authorship and credibility were presence of citations, date of citations, name recognition of the author, presence of bias in the citations, availability of the author for contact, and more. These characteristics were combined to encompass "credibility," which takes into account citations and links as opposed to just the authority of the author.

Feasibility. Characteristics consistent with feasibility need for administrative support, need for consultative support, the amount of time it would take to implement the recommendations provided in the resource, and more. These characteristics were combined to encompass "feasibility" which includes the practicality associated with implementing the recommendations provided in the resource.

Preliminary Item Development

Next, items were generated relative to this literature review. Development and formatting were largely modeled after those employed when developing other URP assessments (e.g., Chafouleas et al., 2009). Initially, 112 items were developed based on items found in previous resources including the URPs as well as Schrock, and others. After 42 items were eliminated after a review for redundancy, 70 items remained for initial content validation activities.

URP-WR Content Validation

Preliminary URP-WR Construction. Four key components of web-resource evaluation arose: accessibility, appearance, credibility, and feasibility. Initial content validation of a hypothesized four factor structure was conducted in the form of a consensus-building task (Hennessy et al., 2016). The author constructed definitions for these four theoretical components using the limited web-based resource literature (e.g., Schrock, 2019). To test these theoretical groupings, nine UCR students and faculty individually rated the 70 items on a scale from 1-4 (1 indicating best fit) indicating fit within the four hypothesized factors. For example, for the item "this resource was easy to access" would hypothetically show best fit with "accessibility," so the participant would mark 1 for accessibility on this item, 2 for potentially appearance, and so on. Given their prior knowledge and experience in assessment, research, and technology utilization, these participants were considered appropriate judges to sort items in a logical or informed rather than random manner. Analysis of rater consensus was used to inform further item reduction and organization of items for pilot administration of the URP-WR.

URP-WR Factor Analysis and Item (Data) Reduction. Once developed, a pilot study was conducted using the drafted URP-WR. The URP-WR was made accessible online through Qualtrics. The online version included the informed consent form, a video detailing instructions for completion of the URP-WR accompanied by written instructions, a link to the web-resource to be evaluated by participants (see Appendix D for a link to the resource), the URP-WR, and finally the URP-A with instructions.

Participants were recruited through email, social media, in class announcements, and through instructors. Participants were emailed the recruitment script that included a chance to win one of ten $25 Amazon giftcards and a link to the survey. Participants first completed the informed consent form, then viewed the video instructions. To facilitate participation understanding of study instructions, a video was provided to supplement written instructions. Completion of the URP-WR included three steps: (a) conduct a Google search, (b) open a link to a provided web-resource, (c) use the Google search and provided web-resource to rate the statements in the URP-WR.

Due to the nature of the URP-WR including an accessibility component, the participants were asked to use a Google search to find their provided resource with a primer of "search for the Good Behavior Game and look for Evidence Based Intervention Network Resources." They were instructed to take no more than two minutes completing this task. They were then instructed to open a link to the provided web-resource and examine it. This resource can be found in Appendix D. They were instructed to spend no more than five minutes examining the resource. Participants were instructed to then move on to the URP-WR. They were told to answer questions that related to a "Google search"

in relation to the search activity, and answer the rest of the questions in relation to the provided web-based resource.

The draft URP-WR included 55 items organized across four hypothesized factors. Participants rated statements on a 6-point Likert scale from (1 being strongly disagree, 6 being strongly agree). Upon completion of the URP-WR, participants were given the option to continue on to complete the URP-A and earn two extra entries into the giftcard drawing, or to stop after completion of the URP-WR and earn one entry. This method was chosen in order to reduce attrition due to length of the study.

Social Validity. Raters who elected to participate used the URP-A to evaluate the acceptability of the URP-WR as an assessment tool itself (Chafouleas et al., 2012). They were asked to rate statements on a Likert scale from 1-6, consistent with general URP formatting. It was made clear that participants were expected to evaluate the URP-WR and not the web-resource provided. The inclusion of this measure provided a measure of social validity of the URP-WR.

Data Analysis

Development and Content Validations

Initially, the consensus-building task was used to demonstrate content validity (see Appendix A). Participant ratings of category of "best fit" were analyzed for each item. Agreement percentages were calculated to evaluate and organize items across hypothesized usability construct. Items that did not reach at least 75% rater consensus in one usability construct were eliminated. Remaining items were organized by category indicated by consensus for pilot administration activities.

Research Questions 1 & 2

Exploratory factor analysis is useful in measure development as it identifies latent factors underlying observed items. There are a number of ways to determine the number of factors extracted. Typically, this includes multiple regression as applied to the correlations between the latent factors and observed indicators (Kline, 2016). In this approach, the underlying or "latent" factors that correspond to the items will be extracted based on how well they predict scores on observed items. If a user rates one item that fits into the "accessibility" factor highly, they should rate another item in that factor similarly. The higher a factor loading, the larger effect the latent factor has on the observed item. This is important because the URP-WR sets out to measure four factors, these factors should predict scores on the URP-WR items. If this hypothesis is supported, the scoring assumption of the IUA has initial support.

Research Question 1. The first research question sought to determine if an initially hypothesized 4-factor structure will emerge from initial use (will a hypothesized factor structure hold (i.e., research identified 4 likely factors?). The data were first examined to ensure it is appropriate for an exploratory factor analysis. Although assumptions are often excluded from analysis in factor analyses, they are included to ensure comprehensive evaluation. This included testing of the following assumptions: interval or ratio level of measurement, random sampling, relationship between observed variables is linear, and a normal distribution (Suhr, 2006). The first two assumptions were addressed through study methodological or procedural choices. Normality was tested through visual inspection of Q-Q plots produced in R. Linear relations between observed

variables was tested through visual inspection of bivariate scatterplots using the "pairs" function in R.

The correlation matrix was inspected and items that correlated significantly ($r \geq .30$) with only one or two other items were deleted to avoid the derivation of meaningless and unnecessary factors. Additionally, items that showed multicollinearity ($r \geq .80$) with at least three other items were deleted to avoid redundancy. Both of these item deletion processes are at the recommendation of Chafouleas et al. (2009). Finally, a Kaiser-Meyer Olkin Measure of Sampling Adequacy was conducted to ensure that the sample size is suitable to factor analysis based on the number of items. According to Kaiser (1974), a value of 0.90-1.00 is marvelous, 0.80-0.89 is meritorious, 0.70-0.79 is middling, 0.60-0.69 is mediocre, 0.50-0.59 is miserable, and below 0.49 is unacceptable.

Once the data were appropriately cleaned and deemed ready, an exploratory factor analysis (EFA) was performed. The number of factors to extract was decided using the R "psych" package and the "fa.parallel" functions. The "fa.parallel" function provided the number of factors to be extracted as determined by a parallel analysis as well as a scree plot. Parallel analysis determines the number of factors to be extracted by extracting factors until the eigenvalues of the real data are less than the eigenvalues of a random data set with the same number of participants (Horn, 1965). The "eigen" function was used to determine eigenvalues of the data, and the "parallel" function was used to determine eigenvalues for random data (these are reported in the results). The number of factors were extracted based on the results of parallel analysis and scree plot, as well as interpretability (i.e. face validity) of the factors extracted (Chafouleas et al.,

2009). Parallel analysis is accurate and commonly used for best practice; other, more conservative, methods of factor extraction such as eigenvalue 1 rule or MAP were not selected because of (a) precedent in previous URP development and (b) parallel analysis produces accurate results that are easy to interpret and thus constitute best practice (Costello & Osborne, 2005). Once the appropriate number of factors was chosen, the "fa" function was utilized to determine factor loadings. This EFA used the Ordinary Least Squared "ols" method of factoring as it is known to produce similar results to the commonly used Maximum Likelihood ("ml") method without an assumption of multivariate normal distribution (Briggs & MacCallum, 2003). The rotation chosen was the "oblimin" oblique rotation as it is assumed that there is a correlation between the factors. All methods of rotation as well as factoring were tested despite to ensure comprehensiveness in the investigation. However, the method and rotation chosen based on logic and common use proved to be the ones that best analyzed the data and were retained for final analysis.

Research Question 2. The second research question sought to determine what item combinations are most appropriate to maximize URP-WR utility (What item combinations are most appropriate to maximize URP-WR utility (i.e., valid results using fewest number of items possible)?). Again, factor analytic techniques were used to address this question. After factors were identified, item loadings were examined using a pattern coefficient matrix. Items that have a pattern coefficient of at least 0.45 on their primary factor will be removed, and items that have a pattern coefficient on a secondary factor above 0.30 will be removed to prevent multidimensionality (Chafouleas et al.,

2009). The “rule of thumb” tends to vary by researcher and depends on the sample size with larger sample sizes allowing for smaller factor loadings to be retained (Yong & Pearce, 2013). With a borderline small sample size of 94, 0.45 as a cutoff is appropriate.

Research Question 3

Research question three concerns the reliability of the URP-WR (i.e., Does the URP-WR demonstrate reliability through internal consistency of items within factors?) The internal consistency was tested using Cronbach’s alpha in R. There is no official guidance as seen with the KMO; however, it seems that an alpha value of at least 0.70 is considered acceptable (Taber, 2017). A high alpha value indicates that the items fit together well, and that a participant who rates one item in the category high on the Likert scale is likely to rank another item in the same category high as well (Blunch, 2008).

Research Question 4

The final research question focused on perceptions of the URP-WR usability. Participants completed the URP-A as a measure of social validity for the URP-WR (Chafouleas et al., 2012). The URP-A is a 28-item assessment using a Likert scale from 1-6. 23 items were used, as five (items 5, 7, 12, 15, and 27) did not apply to IUA established for the URP-WR. Ratings among participants who responded were totaled and aggregated across usability domains. While little guidance is provided for interpreting URP-A scores, higher scores are considered more favorable for the URP-A domains. Thus, the goal is an average overall score at or above 92. This goal of 92 would indicate an average rating of 4 out of 6 on Likert scale items. In general, this would mean participants tended to agree with items. A secondary goal is an average rating per domain

using the same criteria (e.g. there are three items in Factor 2 *Understanding*, so the goal score would be 12 and the best score would be 18). A tertiary goal is an average item score of 4 per category.

Results

Research Question 1

The first research question sought to determine if an initially hypothesized 4-factor structure will emerge from initial use. Initially, a consensus-building task was conducted. Fifteen items (items 4, 12, 17, 19, 20, 23, 27, 28, 29, 41, 51, 55, 57, 61, and 63) did not meet the cutoff of 75% agreement on category of best fit. Thus, 55 items remained and were included in the factor analysis.

Data were examined to ensure appropriateness for factor analysis. The assumption of ratio level of measurement was met due to the use of a Likert scale that functions as a continuous variable in analyses. The assumption of random sampling was violated because of the use of a convenience sample. This is common for pilot studies, but still results should be interpreted with caution. The assumption of normality was violated as the data were largely not normally distributed. This was accounted for through use of the OLS method of extraction, which also is suggested because of the small sample size (Briggs & MacCallum, 2003). The assumption of linear relations between observed variables was met as the bivariate scatterplots demonstrated linear relationships.

All items met the criteria of showing significant correlation ($r \geq .30$) with at least two items as well as not showing multicollinearity with at least three other items. Therefore, no items were deleted due to either multicollinearity (redundancy) or a lack of

correlation (unnecessary items). Thus, all items were included in final analysis. Finally, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy demonstrated a meritorious value, KMO = 0.81. Thus, the data were deemed ready and suitable for factor analysis.

The number of factors to extract was chosen based on a parallel analysis, scree plot, as well as interpretability as determined by the researcher. The break in the scree plot demonstrated that between 4 and 6 factors should be extracted. See Figure 3.2 for the scree plot. The parallel analysis supported this and suggested extraction of four factors. The eigenvalues are reported in Table 3.1. These values indicated that 55.22% of the variance in the data were explained using four factors. When additional factors were extracted, they did not have any items meet the decision rules. Therefore, although additional variance is explained through additional factors, four factors were extracted to eliminate redundancy (and because four factors were suggested by the parallel analysis). The Ordinary Least Squared method and an oblique rotation (oblimin) were used. This rotation demonstrated simple structure and was retained.

The results of the EFA demonstrate a root mean square of residuals (RMSR) value that was close to 0 as expected (RMSR = 0.05). Additionally, the EFA fit value (fit based upon off diagonal values) was equal to 0.97. This is value is acceptable as it is over 0.90 and should be close to 1. Therefore, the fit statistics demonstrate that this model shows good fit and factors can be extracted. A three-factor (fit based upon off diagonal values = 0.96, RMSR = 0.06) and five-factor EFA (fit based upon off diagonal values = 0.98, RMSR = 0.05) were also conducted but were determined to not fit the data as well

as the four-factor model. However, the four-factor model was deemed most appropriate as stated.

Figure 3.1

Scree Plot

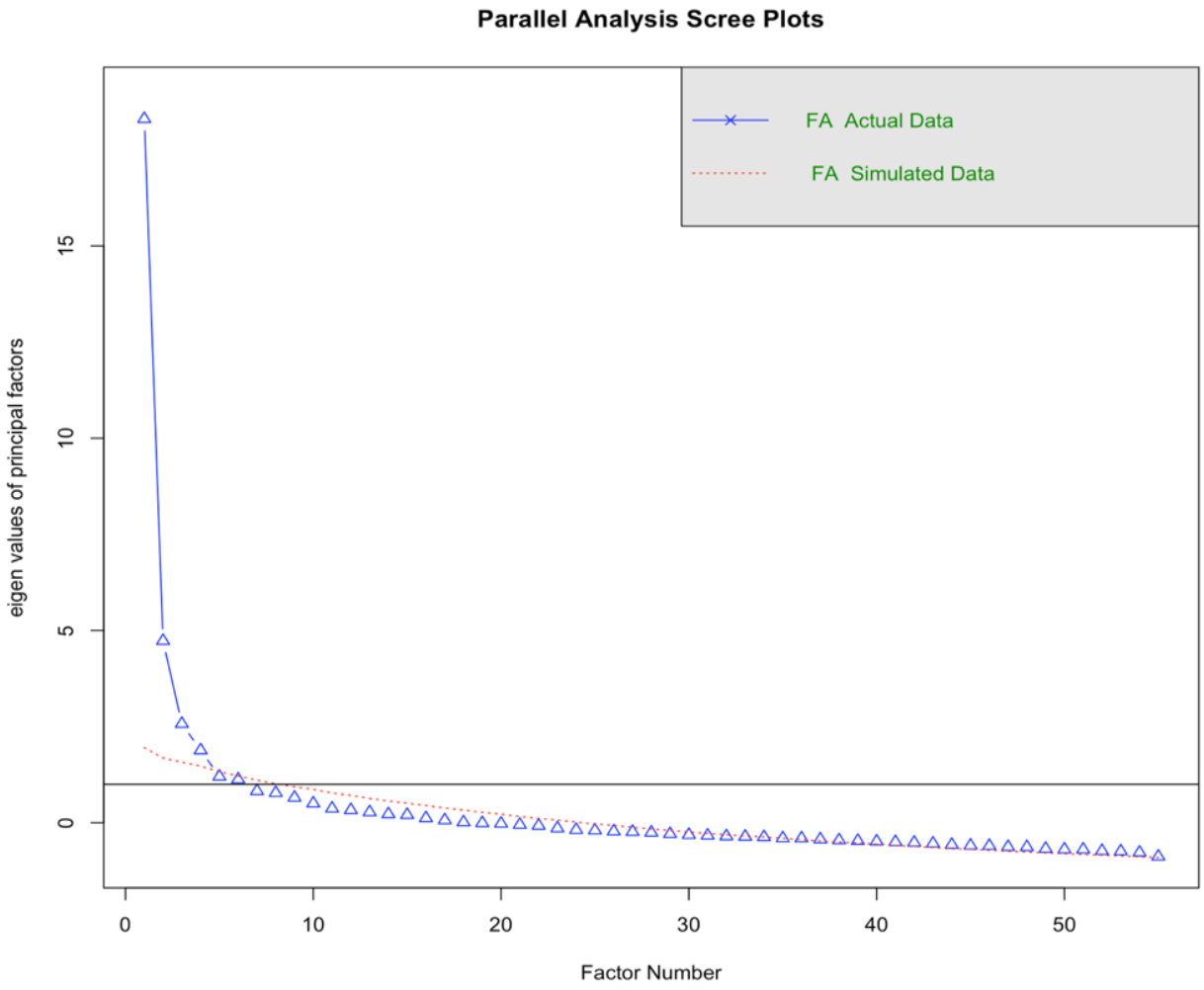

Figure 3.2

EFA Model

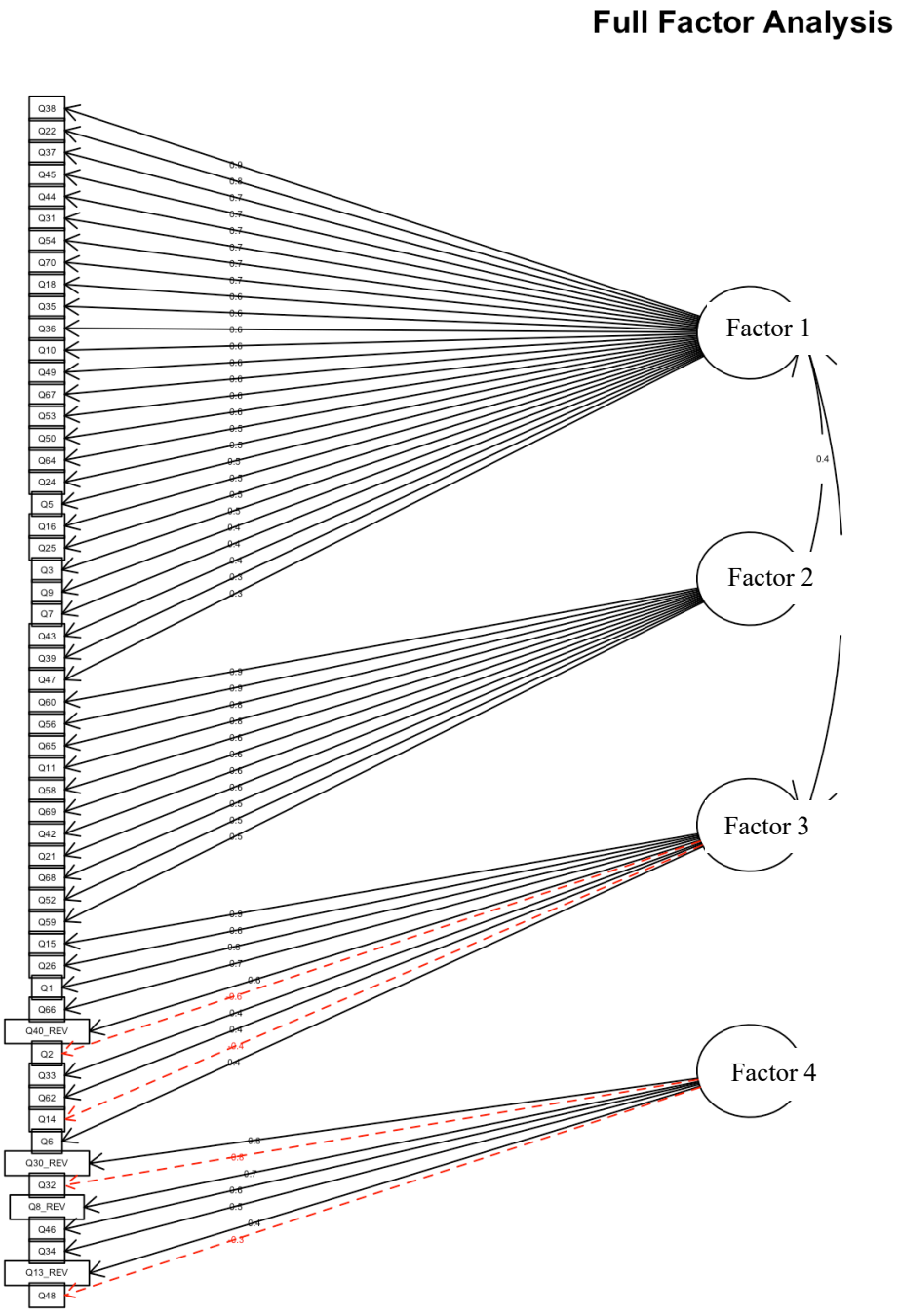

Table 3.1

Eigenvalues and Percentage of Variance Explained

	Actual			Simulated		
Factor	Eigenvalue	% of variance explained	Cumulative %	Eigenvalue	% of variance explained	Cumulative %
1	18.86	34.29	34.29	2.91	4.16	4.16
2	5.53	10.05	44.33	2.71	4.93	9.09
3	3.39	6.16	50.49	2.39	4.35	13.44
4	2.60	4.73	55.22	2.28	4.15	17.59
5	1.98	3.60	58.83	2.39	4.35	21.94

Research Question 2

A second research question sought to determine what item combinations are most appropriate to maximize URP-WR utility. The final items were retained based on factor loading. As precedented in Chafouleas et al. (2009), all retained items had a factor loading of at least 0.45 on their primary factor, and no factor loading above 0.30 on any other factor to avoid multidimensionality. See Table 3.2 for the factor loadings. See Figure 2 for the model of the factor analysis. Based on the decision rules, 34 were retained. Those items are bolded for ease of interpretation. The estimates of communality from the retained items ranged from 0.33 to 0.78.

Table 3.2

Factor Loadings

Item no.	Item	Factor 1		Factor 2		Factor 3		Factor 4		
		P	S	P	S	P	S	P	S	h^2
	Factor 1: Reasonability									
38	The information is from sources known to be reliable.	**0.88**	0.83		0.27		0.13		0.28	0.70
22	The resource provides citations from reliable sources.	**0.79**	0.84		0.35		0.21		0.39	0.71
37	The resource provides citations.	**0.71**	0.71				0.18		0.37	0.53
44	This resource appropriately represents the context of its cited sources.	**0.69**	0.68						-0.27	0.54
45	The sources used by the resource provided appear credible.	**0.68**	0.79		0.30	0.19	0.38		0.42	0.67
54	I understand the components of the recommendations provided in this resource.	**0.65**	0.74						-0.17	0.58
31	I would know what to say if I were asked how to implement the recommendations provided in this resource.	**0.64**	0.67		0.26		0.34		0.22	0.49
70	I believe information from this resource.	**0.63**	0.77	0.27	0.50		0.28		0.34	0.67
18	The resource contains all recommendations needed for implementation.	**0.59**	0.74			0.19				0.57
10	The resource cites its original sources.	**0.59**	0.53			-0.24			0.22	0.33
35	The resource appropriately cites ideas that were not its own.	0.58	0.66		0.30		0.25		0.32	0.46

36	The cost of implementing recommendations in this resource is reasonable.	0.58	0.71	0.28		0.17	0.25	0.21	0.32	0.59
67	I am convinced that the resource is accurate as a direct result of appropriate citations.	0.56	0.75	0.41	0.62		0.20		0.34	0.72
49	I would feel confident sharing these recommendations with my colleagues.	0.55	0.72	0.31	0.51	0.22	0.36		0.31	0.64
53	Information for original resource sources are easily identifiable.	**0.54**	0.60	0.23	0.41		0.17		0.16	0.37
64	I would be willing to implement this resource's recommendations in my setting.	0.51	0.67	0.37	0.55	0.16	0.28		0.22	0.58
5	The information appears to be valid and well-researched.	0.50	0.65	0.33	0.51				0.30	0.51
16	Topics are successfully addressed, with clearly presented arguments and adequate support to substantiate them.	**0.49**	0.61	0.16	0.35		0.17	0.16	0.37	0.42
24	The recommendations could be feasibly implemented in my setting.	0.49	0.64	0.30	0.47	0.25	0.36		0.22	0.53
50	I understand how to implement the recommendations described in this resource.	0.48	0.55		0.16	0.49	0.57			0.52
25	The amount of time required to effectively implement the recommendations provided in this resource is reasonable.	0.46	0.67	0.26	0.43	0.26	0.40		0.40	0.58
9	This resource provides citations in proper APA or another format.	0.46	0.50			-0.25		0.32	0.44	0.38
3	The recommendations provided in this resource could easily be implemented as described.	0.44	0.64		0.28	0.20	0.36	0.29	0.51	0.53
7	I was able to save this resource for future use.	0.40	0.55		0.22		0.16	0.31	0.48	0.39
47	There were no 404 errors or others that blocked me from accessing this resource.	0.31	0.24	-0.23						0.11
43	Quotes and other strong assertions are backed by sources that one could check through other means.	0.39	0.72	0.24	0.39				0.25	0.30

	Factor 2: Appearance									
60	The design of the resource makes me more likely to use it.		0.32	**0.89**	0.88					0.78
56	This resource looks appealing.		0.33	**0.86**	0.86				0.16	0.75
65	I wish more resources were designed the way this one is.		0.37	**0.83**	0.85					0.73
11	This resource is aesthetically pleasing.	-0.16		**0.73**	0.68					0.43
69	The site appears well maintained.	0.29	0.51	**0.66**	0.76				0.18	0.65
58	The resource was updated recently enough for me to trust it.	0.16	0.40	**0.65**	0.71					0.52
42	This resource looks professional.	0.23	0.43	**0.60**	0.69			0.51		
21	Pictures or photographs in the resource add to the information.			**0.57**	0.55	-0.22	-0.25		0.38	
68	This resource looks well organized.	0.36	0.63	0.50	0.64	0.16	0.27		0.34	0.63
59	I was able to download this document as a Word doc or PDF for future use.	0.20	0.32	**0.47**	0.54					0.33
52	There is an image map (large clickable graphic with hyperlinks) on the resource.			**0.47**	0.43	-0.16	-0.22	-0.26	-0.26	0.36
6	This resource was easy to use.	0.34	0.64	0.34	0.50		0.30	0.38	0.57	0.65
39	There are helpful headings and subheadings on the resource.	0.33	0.58	0.30	0.44	0.15	0.29	0.27	0.46	0.49
	Factor 3: System Support									
30	I would need support from my administrator to implement recommendations made in this resource.					**-0.83**	0.78			0.65

32	Support from administration would be needed to implement recommendations provided in this resource.			-0.22	-0.26	**0.79**	0.79		0.20	0.70
8	I could only implement recommendations in this resource with assistance from other adults.					**0.68**	0.65			0.43
46	I could implement the recommendations in this resource by myself.	0.28	0.37			**0.60**	0.65		0.15	0.48
34	I have the skills needed to implement the recommendations provided in this resource.	0.22	0.39			**0.49**	0.57		0.33	0.36
13	Implementation of the recommendations made in this resource would require support from my co-workers.		0.16			**0.41**	0.45	0.25	0.60	0.24
48	This resource ignores important elements from its cited sources.		0.35		0.17	-0.35	-0.43	-0.22	-0.36	0.30
	Factor 4: Accessibility									
15	It was easy to find this resource from a simple Google search.		0.28					**0.84**	0.82	0.68
1	The resource was easy to find.		0.27			-0.17		**0.76**	0.74	0.58
26	It was easy to find this resource.		0.43					**0.80**	0.84	0.73
66	This resource is easily accessible.		0.46	0.21	0.31		0.17	**0.71**	0.77	0.66
2	It was difficult to find this resource from a simple Google search.	-0.28						**0.58**	0.49	0.31
40	This resource required too many links to find.		0.36			0.19	0.34	**0.63**	0.70	0.53
33	It was easy to find this resource without guidance.	0.17	0.41		0.20		0.26	0.44	0.54	0.36
14	This resource took a long time to load.	0.30	0.44				0.23	0.39	0.52	0.35
62	I am satisfied with the amount of time this resource took to load.	0.40	0.59		0.24		0.19	0.42	0.58	0.50

Note. Coefficients below 0.15 were suppressed. Items that meet decision rules are in bold. P = pattern, S = structure.

As anticipated, the factors showed moderate correlation with each other. See Table 3.3 for factor correlations. There was a moderate correlation ($r = 0.35$) between reasonability and appearance, as well as between reasonability and system support ($r = 0.39$). The other factors showed small correlations. For example, accessibility and appearance were not correlated ($r = 0.01$). The moderate correlations necessitated the correctly selected oblique rotation.

Table 3.3

Factor Correlations

Subscale	Factor 1	Factor 2	Factor 3	Factor 4
Reasonability	1.00			
Appearance	0.35	1.00		
System Support	0.24	0.01	1.00	
Accessibility	0.39	0.09	0.21	1.00

Additionally, a second EFA was conducted using the items retained after a reduction of items. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy reached a middling value of 0.79, thus the data were suitable to factor analysis. This EFA also had acceptable fit statistic levels (fit based upon off diagonal values = 0.97, RMSR = 0.06). The four-factor structure emerged once again according to parallel analysis and a scree plot. A three-factor (fit based upon off diagonal values = 0.94, RMSR = 0.08) and five-factor EFA (fit based upon off diagonal values = 0.98, RMSR = 0.05) were also conducted but were determined to not fit the data as well as the four-factor model. The

three-factor model did not capture the complexity of the model with the inclusion of the *system support* factor, and the five-factor model produced a factor that had no items meet the decision rules (similar to the first EFA). This EFA was run to demonstrate that the factor structure still held with the items remaining, and thus is not reported in as much detail as the first EFA. After this EFA, three items (Item 34, 46, and 66) did not meet the decision rule of loading of at least 0.45 on their primary factor and loading no larger than 0.30 on a secondary factor and were thus eliminated.

Finally, the last EFA was conducted on the 31 remaining items. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy reached a middling value of 0.79, thus the data were suitable to factor analysis. The four-factor structure emerged once more according to parallel analysis and a scree plot. The four-factor structure explained 55.77% of the variance in the data. The final EFA had acceptable fit statistic levels (fit based upon off diagonal values = 0.98, RMSR = 0.05). A three-factor (fit based upon off diagonal values = 0.95, RMSR = 0.08) and five-factor EFA (fit based upon off diagonal values = 0.98, RMSR = 0.04) were also conducted but were determined to not fit the data as well as the four-factor model. All items fell within the decision rules, and thus were retained in the final version of the URP-WR. See Figure 3.3 for the scree plot, Figure 3.4 for the final EFA model, Table 3.4 for the Eigenvalues of the final EFA model (used in parallel analysis), Table 3.5 for the factor loadings of the final EFA model, and Table 3.6 for factor correlations of the final EFA model.

Figure 3.3

Final EFA Model Scree Plot

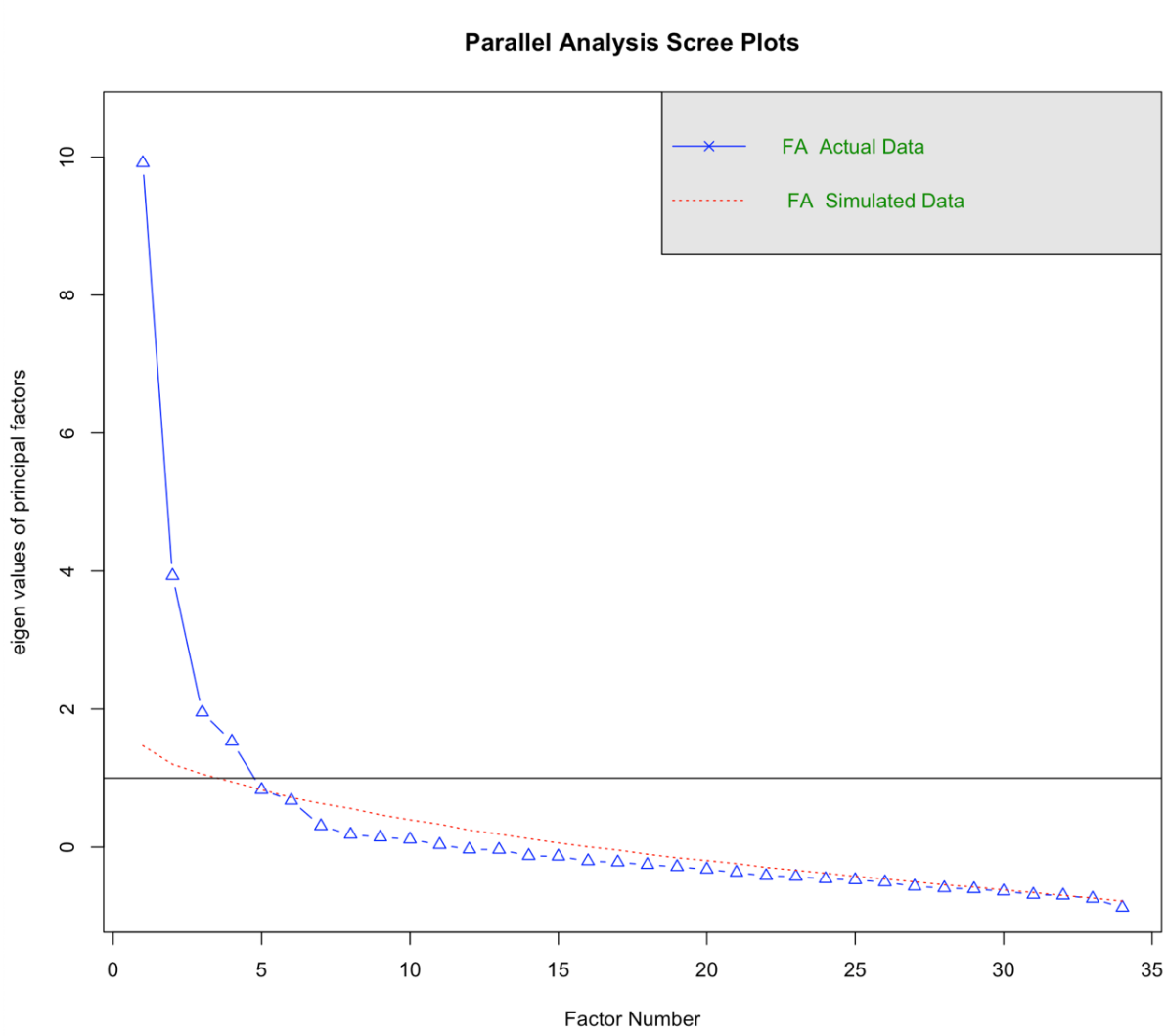

Table 3.4

Eigenvalues and Percentage of Variance Explained Final EFA Model

	Actual			Simulated		
Factor	Eigenvalue	% of variance explained	Cumulative %	Eigenvalue	% of variance explained	Cumulative %
1	10.01	32.29	32.29	2.28	7.35	7.35
2	4.62	14.90	47.19	2.08	6.71	14.06
3	2.66	8.58	50.49	1.93	6.23	20.29
4	2.31	7.45	55.77	1.81	5.84	26.13
5	1.33	4.29	60.06	1.70	5.48	31.61

Figure 3.4

Final EFA Model

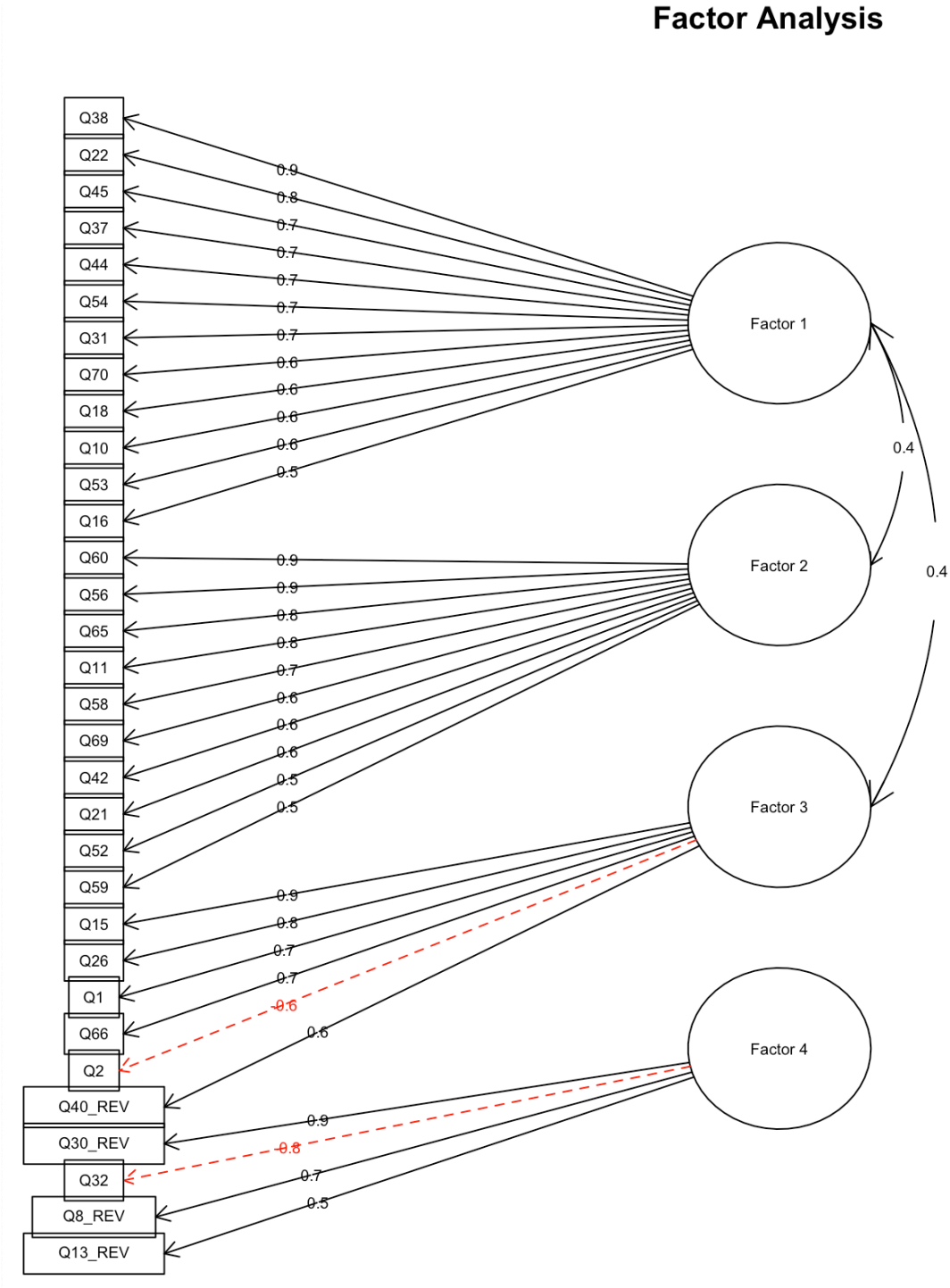

Table 3.5

Factor Loadings for the URP-WR Retained Items

Item no.	Item	Factor 1		Factor 2		Factor 3		Factor 4		h^2
		P	S	P	S	P	S	P	S	
	Factor 1: Reasonability									
38	The information is from sources known to be reliable.	0.91	0.84		0.26				0.24	0.74
22	The resource provides citations from reliable sources.	0.81	0.85		0.34				0.32	0.73
37	The resource provides citations.	0.73	0.71						0.30	0.52
44	This resource appropriately represents the context of its cited sources.	0.70	0.73		0.36				0.22	0.55
45	The sources used by the resource provided appear credible.	0.74	0.80		0.27		0.23		0.35	0.66
18	The resource contains all recommendations needed for implementation.	0.64	0.71		0.30		0.24	0.16	0.30	0.55
54	I understand the components of the recommendations provided in this resource.	0.70	0.74		0.32				0.28	0.56
70	I believe information from this resource.	0.64	0.75	0.24	0.46				0.24	0.61
31	I would know what to say if I were asked how to implement the recommendations provided in this resource.	0.66	0.66		0.23		0.19		0.19	0.45
10	The resource cites its original sources.	0.60	0.55						-0.16	0.33
53	Information for original resource sources are easily identifiable.	0.56	0.62	0.24	0.43		0.16		0.46	0.37

16	Topics are successfully addressed, with clearly presented arguments and adequate support to substantiate them.	0.53	0.64		0.32		0.18	0.31	0.45
	Factor 2: Appearance								
60	The design of the resource makes me more likely to use it.		0.40	0.90	0.88				0.79
56	This resource looks appealing.		0.35	0.86	0.84				0.73
65	I wish more resources were designed the way this one is.		0.40	0.83	0.85				0.73
11	This resource is aesthetically pleasing.		0.19	0.80	0.70				0.51
69	The site appears well maintained.	0.30	0.54	0.63	0.75				0.65
58	The resource was updated recently enough for me to trust it.		0.40	0.66	0.70				0.52
42	This resource looks professional.	0.23	0.46	0.59	0.68				0.51
21	Pictures or photographs in the resource add to the information.			0.56	0.55		-0.23	-0.18	0.34
59	I was able to download this document as a Word doc or PDF for future use.	0.18	0.35	0.48	0.56				0.35
52	There is an image map (large clickable graphic with hyperlinks) on the resource.			0.47	0.46	-0.17	-0.18	-0.19	0.28
	Factor 3: Accessibility								
30	I would need support from my administrator to implement recommendations made in this resource.	0.91				0.91	0.90		0.82
32	Support from administration would be needed to implement recommendations provided in this resource.			0.18	0.28	0.79	0.80	0.18	0.71

8	I could only implement recommendations in this resource with assistance from other adults.			0.75	0.74			0.54
13	Implementation of the recommendations made in this resource would require support from my co-workers.		0.18	0.55	0.59	0.16	0.25	0.38
	Factor 4: Accessibility							
15	It was easy to find this resource from a simple Google search.		0.31		0.59	0.91	0.90	0.77
1	The resource was easy to find.		0.27			0.72	0.72	0.55
26	It was easy to find this resource.		0.44			0.83	0.86	0.77
2	It was difficult to find this resource from a simple Google search.	-0.26			-0.16	0.61	-0.54	0.34
40	This resource required too many links to find.		0.37	0.23	0.32	0.56	0.64	0.48

Note. Coefficients below 0.15 were suppressed. P = pattern, S = structure.

Table 3.6

Factor Correlations for the URP-WR Retained Items

Subscale	Factor I	Factor II	Factor III	Factor IV
Reasonability	1.00			
Appearance	0.36	1.00		
System Support	0.32	-0.10	1.00	
Accessibility	0.13	-0.01	0.17	1.00

Research Question 3

Does the URP-WR demonstrate reliability through internal consistency of items within factors? In order to address this question, estimates of reliability within the four factors were calculated using Cronbach's alpha. All four factors demonstrated acceptable levels of internal reliability as measured by Cronbach's alpha. Factor 1 contained twelve items and demonstrated a high level of internal reliability ($\alpha = 0.93$, 95% CI = 0.90-0.95). These items focused on the citations and believability of the information as well as feasibility of the recommendations of provided in the resource. Thus, this factor was relabeled *reasonability*. A high score on the *reasonability* scale indicates that the user perceived this resource as containing information from credible sources that can be easily understood implemented practically.

Factor 2 contained ten items and demonstrated a high level of internal reliability ($\alpha = 0.90$, 95% CI 0.87-0.93). These items focused on the overall design and appeal of the resource. Thus, this factor was labeled *appearance*. A high score on the *appearance*

scale indicates that the user perceived this resource to be aesthetically pleasing and thus easy to consume.

Factor 3 contained four items and demonstrated an acceptable level of internal reliability ($\alpha = 0.84$, 95% CI 0.79-0.89). These items focused on the overall ease of accessing this resource on the internet. Thus, this factor was labeled *accessibility*. A high score on the *accessibility* scale indicates that the user perceived this resource to be easy to find and without roadblocks to accessibility.

Factor 4 contained five items and demonstrated an acceptable level of internal reliability ($\alpha = 0.82$, 95% CI 0.77-0.88). These items focused on support needed from administration or consultation in order to implement the recommendations provided in the resource. Thus, this factor was labeled *system support*. A high score on the *system support* scale indicates that the user would need a lot of support from the administrators and system in order to implement the recommendations. See Table 3.7 for a breakdown of the reliability statistics.

Table 3.7

Reliability Statistics for Factors

Factor	Average interitem r	SD of interitem r	α	95% CI (α)
Reasonability	0.52	0.01	0.93	0.90-0.95
Appearance	0.49	0.02	0.90	0.87-0.93
System Support	0.57	0.03	0.84	0.79-0.89
Acceptability	0.51	0.03	0.82	0.77-0.88

Research Question 4

Do users perceive this measure as usable? This was assessed through administration of the URP-A. Of the 94 participants that completed the URP-WR pilot study, 75 elected to continue and answer the URP-A questions. The "best" overall score possible on the URP-A would be 138. The goal score was 92. The overall average URP-A score across participants was 98.61. This indicates that the users perceived the URP-WR to be socially valid and acceptable to use in their setting. Additionally, average scores were calculated per category.

The items that fall under the category of *acceptability* are items 1, 9*, 11, 17, 20, 21, and 22 (* indicates reverse coding). Thus, the "best" score for this category would be a 42. The goal score was 28. The average score on this category was a 29.89. The average item score was 4.27. The items that fall into the category of *understanding* are 4, 6, and 24. The "best" score for this category would be an 18. The goal score was 12. The average score on this category was 13.37. The average item score was 4.46. The items that fall under the category *feasibility* are 3, 8, 13, 16, 18*, and 26. The "best" score for this category would be a 36. The goal score was 24. The average score on this category was 27.31. The average item score was 4.55. The items that fall under the category of *system climate* are 10, 14, 19, and 25. The "best" score for this category would be a 24. The goal score was 16. The average score on this category was 16.95. The average item score was 4.24. The items that fall under the category of *system support* are items 2*, 23*, and 28* (reverse coding was chosen to demonstrate the ability to use the URP-WR independently). The "best" score for this category would be 18. The goal score was 12.

The average score on this category was 11.19. The average item score was 3.73. See Table 3.8 for social validity averages.

Table 3.8

Social Validity: URP-A Averages by Category

Category	Average Item Score	Average Score
Acceptability*	4.27	29.89
Understanding*	4.46	13.37
Feasibility*	4.55	27.31
System Climate*	4.24	16.95
System Support	3.73	11.19

Note. *categories that met the goal scores

Discussion

This study sought to provide initial support for and development of the URP-WR. The consensus building task demonstrated initial content validity of the URP-WR. This falls under the scoring inference of the IUA. Participants sorted the items into four categories, and items that did not reach 75% agreement on one category as best fit were eliminated. This resulted in the elimination of 15 items, as well as initial content validation of the URP-WR.

The results of the EFA demonstrated that four factors should be extracted. The EFA demonstrated acceptable levels of fit (0.97) and RMSR (0.06), and thus can be interpreted reliably. This factor structure explained just over half of the variance found in the data. It is satisfactory in social science research to see 50-60% of the variance explained through EFA, thus this is satisfactory (Hair et al., 2010). Subsequent factor

analyses demonstrated similar levels of fit, explanation of variance, as well as the same numbers of factors extracted. However, the four factors extracted were somewhat different than hypothesized. The factors *accessibility* and *appearance* emerged essentially as expected. Items were retained in those factors to measure these aspects of the web-based resource itself. These factors were shown to be important to the evaluation of web-based resources in the literature review. Thus, it makes sense that they emerged separately and significantly in the factor analysis.

Many of the items initially hypothesized to make up the factors of *feasibility* and *credibility* ultimately emerged as one factor. The items related to *credibility* had higher factor loadings than those relating to *feasibility* in general. However, both of these aspects emerged as important to web-based resources in the literature review. Thus, both were included in the final factor that was renamed *reasonability*. This word was chosen as it encompasses the reasonability of implementation as well as the reasonability of the research base supporting the resource. This result was somewhat unexpected, as the research literature did not establish the connection between feasibility of the recommendations and the credibility of them. In fact, this may suggest that the two may be at odds. EBP/I run into roadblocks to due poor implementation fidelity often stemming from feasibility issues. Thus, this result was surprising. It is possible that this connection arose because of the resource used. Participants saw the evaluated WBR as containing credible sources and having feasible recommendations. Thus, the EFA did not pull the two factors apart as expected. This should be empirically evaluated.

The fourth factor *systems support* was not hypothesized, and thus was also surprising. The *systems support* factor also emerged in the factor analysis run by Chafouleas et al. (2009) in the development of the URP-I. Therefore, there is some empirical support for its usage. The *systems support* factor also emerged in the development of the URP-A, but did not reach an acceptable level of reliability (Chafouleas, et al., 2012). This factor did reach an acceptable level of reliability in the current factor analysis. Additionally, this factor contained two items that may warrant reverse coding if the resource is supposed to be used independently. Therefore, interpretability of this factor may be difficult because scoring may need to be reversed in some scenarios but not others. This coupled with the somewhat difficult interpretation of reverse coded items in general is not ideal for the consumer.

The second two factors *accessibility* and *systems support* demonstrated acceptable levels of internal reliability as measured through Cronbach's alpha as hypothesized. The first two factors, *reasonability* and *appearance* demonstrated very high levels of internal reliability as measured through Cronbach's alpha, which is better than hypothesized. Thus, the factors have been shown to measure the same construct through the EFA as well as through a secondary measure.

In terms of social validity, the URP-WR exceeded the goal average score of 92 overall. Additionally, four of the five categories met the goals of an average that was found by multiplying 4 (indicating "agree" on a Likert scale) by the number of items on the scale. This demonstrates that the URP-WR is seen as socially valid, or usable, by the participants (users).

However, there was one category, *system support*, that did not meet the goal. All items in this category were reverse coded to indicate independence in use of the assessment. However, a higher score indicates that the user felt the need for more support from their system to implement the assessment. This does not make the assessment necessarily unusable, just in need of support for use. The use of reverse coded items could affect the scores. A user could mark a higher score than intended due to confusion based on question wording or a preference to agree with the statements (especially if a user is not paying attention) (Suárez-Alvarez et al., 2018). Even then, the category barely missed the goal of an average of 12 (by 0.81). Finally, it should be noted that the goal was decided by the author and not the assessment developer and should be interpreted with caution.

Generally, the URP-WR demonstrated an acceptable level of social validity as defined by the author. The goal scores were met for the overarching scale as well as for four out of five categories measured within the URP-A. Thus, participants viewed the URP-R as acceptable, understandable, feasible, and appropriate to system climate. They also indicated system support would be necessary to carry out the URP-WR. Thus, it may be difficult to implement independently.

Overall, this study demonstrated initial content validity of the URP-WR through the consensus building task, a four-factor structure with 31 items through the EFA, acceptable levels of internal validity within factors, and acceptable levels of social validity in general as well as for a majority of items. Thus, the scoring inference of the URP-WR has been psychometrically evaluated and met.

Implications for Practice

This study began the development and validation of the URP-WR. In order to be used in practice, the URP-WR must undergo further evaluation and validation. Specifically, the implication inference must be tested in order to warrant use in a practical setting. Although it is not yet ready for immediate implementation, the initial development of the URP-WR is promising because it provides a means of evaluating web-based resources that previously went unchecked. Education professionals will be able to use this tool to guide their decision making in order to make appropriate selections in their setting. However, further revisions of the URP-WR should address some key implementation issues. Firstly, the *systems support* factor needs to have very clear instructions regarding interpretation being as the scoring for that item can be confusing. Secondly, feasibility and credibility would ideally arise separately as the literature suggests. Combining the two into one factor is necessary according to the factor analysis but may be confusing to the consumer.

Overall, the URP-WR is still an improvement from currently available tools for web-based resource evaluation. The use of a Likert scale rather than dichotomous items, initial psychometric evaluation, as well as the positive wording found in most of the items makes for a more defensible tool to aid data-based decision making. Thus, the continuing development and improvement of it can result in an important and usable tool for education professionals.

Limitations

The first limitation is that a convenience sample was used in the consensus building task as well as the pilot administration. Thus, there may be characteristics of participants who elect to participate that inherently differ from those that do not or are not selected. This is especially important to take into consideration concerning the social validity assessment as those participants elected to continue after already completing 70 questions. Additionally, the sample of 94 participants is relatively small given the guidelines for exploratory factor analysis. Although some research has shown that sample sizes as small as 50 have produced valid results, the levels of communality extracted from this factor analysis indicate the need for a higher sample size; perhaps such as the 7:1 ratio of items to factors as suggested by Mundfrom et al. (2005). Thus, participants will continue to be recruited to after this analysis for future analyses.

Second, the URP-WR is a self-report instrument and is prone to subjectivity. The perceived quality of a resource may be a different construct than the true quality of that resource for implementation. This assumption will be tested in future validation studies concerning the URP-WR. Third, the URP-WR was used with one intervention central website. This may lead to issues of generalization of the URP-WR to other resources. However, the assumption of generalizability will be tested in future studies concerning the URP-WR.

Finally, some of the estimates of communality (h^2) fell below the suggested value of 0.50. Thus, the interpretation should be taken with caution until replication or future development shows more communality in those items.

Future Directions

Future directions include further validation of the URP-WR. According to Kane's (2013) IUA model, the assumptions underlying an instrument include (a) scoring, (b) generalization, (c) extrapolation, and (d) implication. Further inferences, as well as expanding upon the scoring inference, need to be tested in the validation of the URP-WR. This provides work for future research.

While testing other assumptions, social validity should be reexamined. The social validity measure was used to evaluate the longer, 55 item, draft version of the URP-WR. The length may have impacted user perceptions. The 31-item final URP-WR should be evaluated for social validity again and similarities or differences in results should be noted.

Additionally, future research should address other latent factors underlying a quality web-based resource. These were the four hypothesized based on an initial review of the literature, but there are others that could be implemented. Future research can add additional considerations to the creation of quality web-based educational resources.

Finally, future research should address the connection between credibility and feasibility that emerged. This result was unexpected as credible, evidence-based interventions often run into roadblocks because of poor implementation fidelity. Poor implementation fidelity can emerge from feasibility issues. Therefore, the relationship found between items relating to feasibility and those relating to credibility should be further explored.

Conclusion

Given the increased emphasis on EBP/I use in schools and the increased use of technology to access practice resources, an evaluation of the accessibility, appearance, credibility and feasibility of web-based resources appears both timely and relevant. Development of such an assessment stands to uniquely and substantially contribute to science by creating a much-needed tool that will allow education professionals to evaluate web-based resources. More objective evaluation of resources by these professionals should, in turn support increased and improved implementation of evidence-based practices through use of DBD. A more informed, discerning consumption of web-based resources that are easily accessible, grounded in empirical evidence, and attend to feasibility of implementation has the potential to dramatically increase EBP/I use and subsequent outcomes for students. The creation of the URP-WR may also help bridge the research to practice gap by informing development of new web-based resources and revisions to existing web-based resources. Insights provided by consumers are critical to ensuring the wealth of available EBP/I are disseminated in a manner that is usable, attractive, and accessible for school psychologists and other educators. Furthermore, use of the URP-WR may illuminate strengths and weaknesses of web-based resources. This provides web-based resource developers a manner of evaluating their own resources as they create them. The improvement in quality of resources coupled with the improvement in consumers' ability to evaluate resources will greatly improve the practice of web-based resource use.

References

Andrews, T.M. (2020). Our iPhone weekly screen time reports are through the roof, and people are 'horrified.' *The Washington Post*. Retrieved from: https://www.washingtonpost.com/technology/2020/03/24/screen-time-iphone-coronavirus-quarantine-covid/

Benotsch, E. G., Kalichman, S., & Weinhardt, L. S. (2004). HIV-AIDS patients' evaluation of health information on the internet: the digital divide and vulnerability to fraudulent claims. *Journal of consulting and clinical psychology*, *72*(6), 1004.

Blunch, N. J. (2008). Introduction to structural equation modelling using SPSS and AMOS. Thousand Oaks, CA: Sage Publications Ltd.

Briesch, A. M., Chafouleas, S. M., Neugebauer, S. R., & Riley-Tillman, T. C. (2013). Assessing influences on intervention use: Revision of the Usage Rating Profile-Intervention. Journal of School Psychology, 51, 81–96.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*(1), 25–56.

Buchanan J., Kock N. (2001) Information Overload: A decision making perspective. In: Köksalan M., Zionts S. (eds) Multiple Criteria Decision Making in the New Millennium. Lecture Notes in Economics and Mathematical Systems, 507. Springer, Berlin, Heidelberg.

Cattell, R. B. (1978). The scientific use of factor analysis. New York: Plenum

Carnine, D. (1997). Bridging the Research-to-Practice Gap. *Exceptional Children, 63(4),* 513–521.

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., & McCoach, D. B. (2009). Moving beyond assessment of treatment acceptability: An examination of the factor structure of the Usage Rating Profile – Intervention (URP-I). *School Psychology Quarterly***,** *24*, 36-47.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education & Treatment of Children, 34(4),* 575–591.

Chafouleas, S. M., Miller, F. G., Briesch, A. M., Neugebauer, S. R., & Riley-Tillman, T. C. (2012). *Usage Rating Profile – Assessment*. Storrs, CT: University of Connecticut.

Chafouleas, S.M., Briesch, A.M., McCoach, D. B., & Dineen, J.N. (2018). *Usage Rating Profile – NEEDS.* Storrs, CT: University of Connecticut.

Cook, D.A., Brydges, R., Ginsburg, S., Hitala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education, 49,* 560-575.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, *10*(1), 7.

Cummings, J.A. (2011). Technology in the practice of School Psychology: The future is past tense. *Oxford Handbooks Online.*

de Winter, J.C., Dodou, D., Wieringa, P.A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44(2),* 147-81.

Fischer-Baum, R. (2017). What 'Tech World' did you grow up in? *The Washington Post*. Retrieved from: https://www.washingtonpost.com/graphics/2017/entertainment/tech-generations/?noredirect=on

Greenwald, H.J., & O'Connell, S.M. (1970). Comparison of dichotomous and Likert formats. *Psychological Reports, 27,* 481-482.

Greenwood, C. R., & Abbott, M. (2001). The research to practice gap in special education. *Teacher Education and Special Education, 24(4),* 276–289.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate Data Analysis (7th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Hennessy, S., Rojas-Drummond, S., Higham, R., Marquez, Ana María, Maine, F., Ríos, R.M., García-Carrión, R., Torreblanca, O., Barrera, M.J. (2016). Developing a coding scheme for analyzing classroom dialogue across educational contexts. *Learning, Culture, and Social Interaction, 9*, 16-44.

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. *Best Practices in School Psychology, 4*, 993–1006.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30(2)*, 179-185.

Huberman, M. (1994). Research utilization: The state of the art. *Knowledge and Policy 7*, 13–33.

Johnson, A., Ledoux, M., Bains, B., Maggin, D., Buren, M., Couch, L. (2019, Feb., 28). *Special educator research utilization: Bridging the gap* [Conference Session]. NASP 2019/Hyatt Regency, Atlanta, Atlanta, GA, United States.

Kaiser, H. F. (1974). An Index of Factorial Simplicity. *Psychometrika, 39*(1), 31--36.

Kamphaus, R.W. & Frick, P.J. (2005). *Clinical Assessment of Child and Adolescent Personality and Behavior*. New York: Springer.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1),* 1-73.

Kerr, M.M. & Nelson, C.M. (2006). *Strategies for Addressing Behavior Problems in the Classroom (5th ed).* Upper Saddle River NJ: Pearson Education.

Kline, P. (1994). An easy guide to factor analysis. New York: Routledge.

Koeze, E., & Popper, N. (2020). The virus changed the way we internet. *The New York Times.* Retrieved from: https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html

Kratochwill, T. R., & Shernoff, E. S. (2003). Evidence-based practice: promoting evidence-based interventions in school psychology. *School Psychology Quarterly, 18*(4), 389-408.

Lydia M. Olson Library (2018). Evaluating Internet Resources. Retrieved from: https://lib.nmu.edu/help/resource-guides/subject-guide/evaluating-internet-sources

Miller, F.G., Neugebauer, S.R., Chafouleas, S.M., Briesch, A.M., Riley-Tillman, T.C. (2013). *Examining Innovation Usage: Construct Validation of the Usage Rating Profile-Assessment* [Poster Session]. APA 2013/Hawai'i Convention Center, Honolulu, HI, United States. Retrieved from: https://dbr.education.uconn.edu/wp-content/uploads/sites/916/2015/07/2013-APA-Miller-Neugebauer-Chafouleas-Briesch-Riley-Tillman.pdf

Miller, M. (2012). 53% of organic search clicks go to first link. *Search Engine Watch*. Retrieved from: https://www.searchenginewatch.com/2012/10/10/53-of-organic-search-clicks-go-to-first-link-study/

Morr, S., Shanti, N., Carrer, A., Kubeck, J., & Gerling, M.C. (2010). Quality of information concerning cervical disc herniation on the internet. *The Spine Journal, 10(4)*, 350-354.

Mundfrom, D.J., Shaw, D.G., & Lu Ke, T. (2005) Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5:2*, 159-168.

Noell, G. H., & Gansle, K. A. (2014). Research examining the relationships between consultation procedures, treatment integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), Handbook of research in school consultation (2nd ed.; pp. 386-408). New York, NY: Taylor & Francis Group/Routledge.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

Ream, E., Blows, E., Scanlon, K., & Richardson, A. (2009). An investigation on the quality of breast cancer information provided on the internet by voluntary organisations in Great Britain. *Patient education and counseling, 76(1), 10-15.*

Ribble, M. (2012). Digital citizenship for educational change. *Kappa Delta Pi Record, 48(4),* 148-151.

Schoenwald, S.K., Garland, A.F., Chapman, J.E., Frazier, S.L., Sheidow, A.J., Southam-Gerow, M.A. (2011). Towards the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38,* 32-43.

Schrock, K. (2019). Schrock Guide. Retrieved from: https://www.schrockguide.net/critical-evaluation.html

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and treatment of children*, 351-380.

St. Peter Pipkin, C.C., Vollmer, T.R., & Sloman, K.N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translational model. *Journal of Applied Behavior Analysis*, *43,* 47-70.

Sterling-Turner, H. E., & Watson, T. S. (2002). An analog investigation of the relationship between treatment acceptability and treatment integrity. *Journal of Behavioral Education, 11*, 39–50.

Suárez-Alvarez, J., Pedrosa, I., Lozano, L.M., García-Cueto, E., Cuesta, M., Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30(2),* 149-158.

Suhr, D. (2006). Exploratory or confirmatory factor analysis. SUGI 31 Proceedings. Retrieved from: https://support.sas.com/resources/papers/proceedings/proceedings/sugi31/toc.html

Taber, K.S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273-1296.

Thurstone, L. L. (1947). Multiple factor analysis: A development and expansion of vectors of the mind. Chicago: University of Chicago.

Witt, J. C., & Martens, B. K. (1983). Assessing the acceptability of behavioral interventions used in classrooms. Psychology in the Schools, 20, 510–517.

Xian, S., Hia, H., Yin, Y., Zhai, Z., & Shang, Y. (2016). Principal component clustering approach to teaching quality discriminant analysis. *Cogent Education, 3(1)*, 1194553.

Yong, A.G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9(2)*, 79-94.

Appendices

Appendix A - Consensus Building Task

https://docs.google.com/document/d/1lYSDGmJaHLQLgKDA08oXa9tFYwqZZtGHZsKa03Z2v_0/edit?usp=sharing

Appendix B - URP-WR Pilot Form

https://docs.google.com/document/d/1BHOM2UU7_El3Yd9Pposi7KIBYz4nk2redP1UTg6JEkw/edit?usp=sharing

Appendix C - URP-A
https://production.wordpress.uconn.edu/educationurp/wp-content/uploads/sites/965/2014/09/URP-A.pdf

Appendix D - Web-Based Resource

https://www.interventioncentral.org/behavioral-interventions/schoolwide-classroommgmt/good-behavior-game

Appendix E - Hypothesized Item Factors

https://docs.google.com/document/d/1d8FAgPPtyjia4uggJCvat2fElVf821L8WI7YBTNowlU/edit?usp=sharing

Appendix F - Actual Consensus Building Task Results by Item

https://docs.google.com/document/d/1g-herscMuUkO0VkfFyobVPGrLTx593-f-HahSmH6y5s/edit?usp=sharing

Appendix G - Final Retained Items URP-WR

https://docs.google.com/document/d/1RN03hgqL3CwAuAnvXfhoIzsO1Mr85gEf3W_uitrJ9wc/edit?usp=sharing