Title

Examining the Validity of Inferences about Intervention Implementation Based on the Usage Rating Profile-Web Resource

Author
Mandracchia, Nina Rosalie

Publication Date
2022

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Examining the Validity of Inferences about Intervention Implementation Based on the Usage Rating Profile-Web Resource

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

Doctor of Philosophy

in

Education

by

Nina Rosalie Mandracchia

June 2023

Dissertation Committee:
Dr. Austin Johnson, Chairperson
Dr. Stephanie Moore
Dr. Marsha Ing

The Dissertation of Nina Rosalie Mandracchia is approved:

____

____

____

Committee Chairperson

University of California, Riverside

Acknowledgment

Although not everyone who contributed to the success of this dissertation can be acknowledged in just one page, I would like to acknowledge and sincerely thank several incredible people. First, I would like to thank my committee chair Dr. Austin Johnson for his supervision throughout my years in the program. He has been a wonderful advisor, program director, and professor throughout my years at UCR. I would also like to thank my other committee members Dr. Marsha Ing and Dr. Stephanie Moore. Dr. Ing's class on advanced psychological test and measurement inspired the idea behind this dissertation, and her guidance throughout this process has been irreplaceable.

I would also like to thank my internship supervisor Dr. Kavita Atwal. Her guidance, flexibility, and support have allowed me to balance my dissertation and my internship in a way that I feel lucky to have experienced. Next, my cohort deserves many thanks for the years of study groups, tacos, and friendship. Jessica, Manasi, Michelle, Theresa, Tyler, and Vanessa, thank you for everything.

I cannot begin to express enough gratitude for my family. To my parents, Sal and Jenny Mandracchia. Thank you for providing the best parental support I could have asked for. From offering to buy me a new laptop when this one overheated from running too many bar graphs, to making sure that I remembered to eat, I could not have done this, any of this, without your support. Last but not least, to my fiancé Phil. You are the motivation I needed to complete this degree. I cannot wait to be your "Dr. Nina." Thank you for finding me and showing me what it means to have and to be a truly loving and supportive partner.

ABSTRACT OF THE DISSERTATION

Examining the Validity of Inferences about Intervention Implementation Based on the
Usage Rating Profile-Web Resource

by

Nina Rosalie Mandracchia

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, June 2023
Dr. Austin Johnson, Chairperson

There are a range of educational web-based resources available for use by education professionals. Although widely available, such web resources vary in terms of usability. As of 2022, there was no way to assess whether these web resources are useful for educators in terms of identifying new classroom interventions to implement. In response, the Usage Rating Profile-Web Resource (URP-WR) w as designed to measure the usability of educational web resources. Given this intended use of the URP-WR, this dissertation evaluates how potential users actually use information from the URP-WR to make decisions about an intervention web resource. Two groups of potential users were included in this study: pre-service teachers (n = 37) a nd doctoral students (n = 33) t o allow for comparisons of decisions of use. First, participants were asked if they would be willing to implement a new intervention designed to improve their teaching. Participants were then asked to rank the factors that they prioritized from 1-5 with 1 being most influential and 5 being least influential to their decision making. Participants were provided these questions five times accompanying five presentations of data (i.e., Scenarios). The five scenarios were designed to represent five possible website types that participants may encounter: (a) high on all five factors of the URP-WR, (b) low on all five factors, (c) medium on all five factors, (d) high on credibility, low on appearance, accessibility, feasibility, and system support (e) low on credibility, high on appearance, accessibility, feasibility, and system support. Participants were asked to respond narratively to the question “Thinking back to all five scenarios, why did you rank the factors in the way that you did?” Results suggest that pre-service teachers are more likely than doctoral students to endorse intervention uptake in situations where data do not support usefulness of factors as well as when data support all factors but credibility. Participants largely agreed that credibility information from the URP-WR was most influential to their decision to use the web resource intervention, while appearance and system support were least influential. Implications, limitations, and future directions of the URP-WR are discussed.

Table of Contents

CHAPTER 1

CHAPTER 2

CHAPTER 3

List of Figures and Tables

Introduction

Digital technology is inseparable from the social and professional lives of most human beings. Over five years ago in 2016, 88% of households had access to the internet (Fischer-Baum, 2017), and in 2019 over 5 billion people worldwide owned a smart phone (Silver, 2019). Reliance on internet is even greater after the COVID-19 pandemic, with worldwide internet traffic increasing 15-20% within the first week of the pandemic and only growing since (Feldman et al., 2021). Given near-ubiquitous access to digital resources, it is critical for evidence-based information to show relevance to modern digital expectations, particularly when it comes to education. However, educationally related information available on the internet comes in varying qualities; some resources provide invaluable information, while others fall short (Beahm et al., in submission; Mandracchia & Sims, 2020).

Educational technology, although historically slow to keep up with technological expectations (Koehler & Mishra, 2005), has firmly established space among the universe of resources for educational professionals. For instance, the communication and classroom activity app ClassDojo claims to be used in two-thirds of schools in the US and has expanded to 180 countries (ClassDojo, n.d.). The web-based marketplace TeachersPayTeachers has an active community of 7 million users and claims that 85% of PreK-12 US educators use their site (TeachersPayTeachers, n.d.). Despite their uptake, there is limited information available on the extent to which these resources reflect evidence-based practices; one study has gone so far as to suggest that the content model of TeachersPayTeachers "implicitly redefines what constitutes an education, elevating holiday activities and classroom décor to the same level as established curriculum" (Shelton et al., 2021, p. 1). In their review, Shelton et al. (2021) found that only 11% of resources reviewed aligned with learning standards, and that nearly half of all resources had a 4-star rating and nearly half had a 0-star rating, rendering the rating system all but irrelevant. On the other hand, evidence-based educational resources are also available online (e.g., What Works Clearinghouse; U.S. Department of Education, n.d.), although a quick glance at either website demonstrates the gap between their aesthetic qualities and those of more widely used, for-profit websites.

As educational practices evolve and more research is conducted, it is incumbent upon educators and educational professionals to stay knowledgeable of such changes. Traditionally, such continuing education has taken the form of educational professional development (Avalos, 2011), although teachers often perceive these as demonstrating inconsistent quality or promoting impractical recommendations (Borko, 2004; Borko et al., 1997; Putnam & Borko, 2000). In the absence of relevant activities, or if these activities do not provide sufficient information using effective strategies, educators with questions regarding what to do often turn to a trusted colleague or the internet (Buren et al., 2021).

Little is known about the decisions that educators make using information gleaned from web resources they find. It is well established that teachers and education professionals, especially those younger and more inexperienced in the field, use web resources (Buren et al., 2021; Opfer et al., 2016), but what is not yet known is what decisions educators make resulting from these resources. Decisions about using these web resources depend largely on personal characteristics and previous beliefs and thus can be quite variable (Coburn et al., 2009a; Coburn & Turner, 2011). In addition, the inferences made by different populations of interest (e.g., pre-service or in-service teachers, doctoral students, faculty researchers in education) may also vary due to differing preexisting worldviews and other factors affecting decision-making including type and level of training (Coburn et al., 2009b; Moss, 2016).

Decisions about using web resources may also be influenced by data regarding other users' experiences with these resources (see Fogel & Zachariah, 2017; Luca, 2016). For example, the website Yelp provides a way for users to share their experiences about restaurants with other users. Other users read the reviews on Yelp and ostensibly use those data to make decisions about whether to try a new restaurant. Luca (2016) found that Yelp reviews had a significant impact on the restaurant industry, suggesting that Yelp may have taken over traditional forms of reputation. Fogel and Zachariah (2017) found that increased brand trust and increased number of reviews read were predictive of intentions and behaviors regarding the product being reviewed, indicating that reviews can lead to purchase of a product. Such research has potential relevance to the adoption of an intervention in the context of the Usage Rating Profile-Web Resource (URP-WR). The URP-WR provides one potential way for users to share their experiences around educational resources, which could influence other users' intervention uptake similarly to how Yelp influences restaurant choice or product purchase. Although the URP-WR was designed for this intended use, there is no evidence that users actually find it useful in making decisions about implementing an intervention.

This dissertation aims to explore variation in inferences based on data derived from a measure of educational web resource usability, the URP-WR, from two populations of interest: educational researchers in training and pre-service teachers. In doing so, this study begins to address the validity of inferences of the URP-WR. If inferences vary between these two populations of interest, there may be subsequent concerns as to the current usefulness of the URP-WR for making decisions about intervention uptake, as well as necessary considerations for future use.

Review of Relevant Literature

Internet Information Quality

The current quality of educational information available on the internet varies (see Beahm et al., in submission; Test et al., 2015). Development services like WordPress and Weebly combined with easily accessible web-page hosting services like GoDaddy make webpage setup very easy for developers (WordPress, n.d.). These services have substantial benefits including the potential for the equitable dissemination of quality information to audiences who would typically not have access (Lindsay & Poindexter, 2003).

However, one potential drawback includes the increased dissemination of poor-quality information (Polikoff, 2019; Shelton et al., 2021; Test et al., 2015). For instance, since medical patients often use internet searches as a resource that does not require insurance or out-of-pocket expenses, researchers have investigated the quality of web-based medical information patients are accessing. Researchers found that medical resources on the internet provided low quality information on HIV/AIDS (Benotsch et al., 2004), breast cancer (Ream et al., 2009), and cervical disk herniation (Morr et al., 2010). In the field of education, only one empirical peer-reviewed exploration has been identified, perhaps in part due to the lack of a measure of web resource usability. Test et al. (2015) awarded only 34% of 47 educational websites that claimed to promote evidence-based practices a designation of "trust." In another project, expert reviewers indicated that 64% of resources evaluated from TeachersPayTeachers, ReadWriteThink, and Share My Lesson "should not be used" with a majority of pages on all three sites being rated a 0 or a 1 on a 0-3 quality scale (Polikoff, 2019).

Given the ease of dissemination and information quality concerns generally, the presence of low-quality educational information in web-based resources in the field of education seems plausible, if not likely, with preliminary research providing support for such a contention (i.e., Test et al., 2015; Polikoff, 2018). This may be especially relevant for subjects that are not well understood in practice. For example, the concept of "learning styles" is a neuromyth (i.e., a practice that claims to be founded in neuroscience but is in actuality derived from a misconception of a neuroscientific result; Howard-Jones, 2014) that is nonetheless commonly believed and utilized in the field of education. Newton and Salvi (2020) conducted a review of 37 studies spanning across years and across the globe and found that self-reported belief in matching instruction to learning styles (89%) and self-reported use of matching instruction to learning styles (80%) was high among teachers, with no sign of deterioration as 95% of pre-service teachers reported belief in learning style instruction matching. Resources on learning styles abound across the internet (e.g., Malvik, 2020; Education Planner, n.d.), perpetuating neuromyths as seen in Newton and Salvi (2020) and thereby providing information that has no empirical support and may hinder student learning.

Research to Practice Gap

Even when high-quality information is available, it can be difficult for practitioners to use this information effectively in practice. The “research to practice gap” refers to the lack of quality conversation between researchers and practitioners in the field (Carnine, 1997). This is not a new concept, and it goes by different names (e.g., the “bench to bedside gap” in the medical field; Wolf, 1974), but one core component of this gap is attributable to the implications and feasibility of research-based practices. Research may identify a practice that produces significant results in a clinical or contrived setting, but sometimes the intervention needed to produce that result is lacking the feasibility needed to transfer to the “bedside” or practical setting (Carnine, 1997). Additionally, the implications of the research may not be significant enough in the eyes of the practitioner to warrant the additional time and effort spent in completing the intervention. Indeed, the diffusion of innovations and adoption of interventions has developed into an entire interdisciplinary field of research in the form of implementation science (e.g., Sanetti & Luh, 2019). Compounding these challenges, practitioners often simply lack access to evidence-based practices; for example, special educators have reported a variety of issues in accessing evidence-based practices including terminology changing, not knowing where to look, and lack of access to scholarly journals (Buren et al., 2021). Furthermore, peer-reviewed journal access is expensive and uses jargon that may not be familiar to those outside of that particular research sphere.

To build a bridge between research and practice, high-quality information must be available and detectable, and decisions regarding use that are made based on these resources must be understood. The internet provides plenty of high-quality and usable information. However, such high-quality information can be difficult to find due to a lack of training in high-quality web-resource identification, differences in data use that have not been explored or understood, and a lack of an appropriate evaluation tool.

Current Web Resource Evaluation Tools

There is a dearth of measures specifically designed to evaluate web resources; those that are available tend to be in the forms of guidelines provided by university library websites (see Lydia M. Olson Library, 2018). There are even fewer tools designed specifically for educators, although educator Kathy Schrock has a website dedicated to individually-developed checklists for educators evaluating web resources (Schrock, 2020).

Such library-associated tools provide frameworks for users to ask themselves questions related to criteria determined to comprise a quality resource. For example, the Lydia M. Olson Library (2018) website through Northern Michigan University has a page entitled “Evaluating Internet Sources.” This page provides six criteria: authority, accuracy, objectivity, currency, coverage, and appearance. In working through this web-based tool, users are prompted to ask themselves certain questions when reviewing a web resource. Some examples include: “Is it clear who is responsible for the contents of the page?” or “Does the content appear to contain any evidence of bias?” Although useful as a heuristic for individual critical thinking, this does not provide quantitative information on the resource itself and lacks extant research regarding the quality of the resulting data, thereby limiting its usefulness for broader evaluative purposes.

Schrock (2020) provides an “ABCs” of website evaluation with criteria similar to those of the aforementioned university libraries, including authority, efficiency, and verifiability. On the individual tools, questions are accompanied by dichotomous (yes/no) response choices. Some questions on these tools are: “Does the page take a long time to load?”, “Does the information appear biased?”, and “Are the facts on the page what you were looking for?” Educators using this tool are expected to construct a narrative summary addressing whether the website was helpful after completing the series of about 10-20 dichotomous questions. The narrative summary then provides the basis for comparison between websites, although guidance or steps for how to make appropriate comparisons (i.e., a website that the user characterizes as trustworthy is “better” than one the user characterizes as “easy to use”) is not provided.

Although these evaluations provide information to consumers of web-based resources, they have two key limitations. The first is that the data derived from these measures cannot be easily quantified for use by other practitioners. In other words, the user must fill out or answer the items for each resource they encounter instead of having an average rating from numerous other professionals in their field. This is not a fault of the measure, as the organizations and individuals that created these evaluation tools did not appear to aim to provide aggregate data. Nevertheless, educators who are often overworked and under pressure to produce quick and meaningful results (e.g., Alson, 2019; Hester et al., 2020) may find this type of evaluation tool infeasible to use frequently.

Another key limitation is that these tools have not been empirically evaluated. In other words, assumptions that underlie measures and data interpretation have not been checked (Kane, 2013). There is no data to support the alignment between the intended assumptions of how the data should be used and the actual data use (Ing et al., 2021). Although it may not be necessarily problematic if the intended and actual data use practices do not align (Ing et al., 2021), it is problematic that the actual data use is not understood, as poor and harmful data-use practices have no way of being identified.

Evidence-Based Intervention Identification

Fortunately, there are numerous evidence-based programs available to promote positive outcomes for students across a wide variety of student outcomes. For example, Merrell's Strong Kid series (Merrell, 2008) is a K-12 social-emotional curriculum that has demonstrated effectiveness of decreased internalizing and externalizing behaviors, increased prosocial behaviors, retention of content knowledge, and effective cultural adaptations (e.g., Cramer & Castro-Olivo, 2015; Gueldner & Merrell, 2011; Kramer et al., 2014; Marchant et al., 2010). These effects were amplified when combined with an evidence-based multitiered system of support known as Positive Behavioral Intervention and Support (PBIS; Cook et al., 2015).

Evidence-based interventions to promote positive academic outcomes are similarly plentiful. For instance, reciprocal teaching (RT; Palincsar & Brown, 1986) is a well-known evidence-based practice developed in the late 1980s to improve students' reading comprehension. RT has been effective across cultures (Tarchi & Pinto, 2016) and grade levels (Okkinga et al., 2018), even extending to online learning (Yang, 2010). It has continued to accumulate evidence through recent years (Hamdani, 2020; Pilten, 2016).

Although these interventions exist, teachers lament not being able to find resources to aid their implementation due to accessibility roadblocks resulting from the research-to-practice gap (Buren et al., 2021). Teachers report using the internet for intervention/lesson-plan identification more so than any other use, including social networking (Choi et al., 2018). Resources such as What Works Clearinghouse (WWC; US Dept. of Education, n.d.) provide evaluations of intervention quality based on evidence which can be helpful to teachers. However, WWC requires some considerable preliminary effort from users, in that teachers first need to identify an intervention, find it on WWC, interpret the findings of WWC (if the intervention exists on WWC at all), and then find tools to aid implementation. A tool evaluating usability of web resources as teachers find them which includes an abbreviated evaluation of evidence has the potential to cut multiple steps out of this process. It also adds other dimensions of usability (such as accessibility of the web resource or feasibility of the recommendations) that sources like WWC do not currently cover. Therefore, a measure that aids teachers in the evaluation of available web resources is worthy of development, especially one that is developed under contemporary understandings of validity.

Interpretation/Use Argument Approach to Validity

The development of a new measure should investigate validity, or “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA, APA, NCME, 2014, p. 11). One heavily cited method for evaluating the validity of inferences is the Interpretation/Use Argument (IUA) approach proposed by Kane (1992; 2013a; 2013b; 2016). The first step of the IUA approach is to clearly state the proposed use and interpretation of the measure. This provides potential users with a clear understanding of what they should use the measure for, when they should use the measure, and how to interpret the results of the measure. After the IUA has been identified, the developer conducts a series of experiments designed to test whether the measure can and should be used as intended.

Kane (2013b) identified four areas for testing the IUA: (a) scoring (providing a correct, replicable, quantifiable score to an observation), (b) generalization (gathering a representative sample), (c) extrapolation (applying the scores to real life scenarios), and (d) implication (using that data to make a decision). The IUA validation process requires that any inferences (or assumptions) be checked empirically in order to support interpretation. This is an arduous task; therefore, continuous validation and revision of a measure constitute best practice. Although the IUA approach provides a unified framework for understanding the properties of data derived from an instrument, as well as the use of those data, it has been rarely utilized in the educational sciences.

One example of a validation process in educational measurement is that of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS; Kilgus et al., 2013;

2015). In terms of a model for interpretation for the SAEBRS, an exploratory factor analysis suggested a three-factor model consisting of *social behavior*, *academic behavior*, and *general behavior* (Kilgus et al., 2013). For model for use, the cut scores established by Kilgus et al. (2015) indicate whether students should be placed into one of two categories for each of the three factors: at risk or not at risk. The development process employed across these studies examines the internal structure of the measure, which is in line with one portion of Kane's IUA (i.e., scoring). However, it also demonstrates how educational measurement research has largely not yet extended to the consequences of actual use practices (i.e., implication), which play a key factor in the IUA validation process (Cizek, 2016; Haertel, 2013; Shepard, 2016).

Evidence of Actual Interpretations and Uses

Consensus regarding the accumulation of validity evidence has shifted from supporting the overall "validity of the measure" to supporting validity of use of a measure in specific contexts and for specific interpretations and uses (AERA, APA, & NCME, 2014). As a result of this shift, there is a need to gather evidence in actual contexts and measure inferences that real users make, instead of solely those that test developers intend them to make (Ing et al., 2021). In other words, it cannot be assumed that users will in reality use measures in a manner intended by a developer, or that they will make use of data in a manner intended by a developer; therefore, this needs to be empirically evaluated. In addition, different users bring their own perspectives and values to the interpretation of the measure (Moss, 2013; Moss 2016) which further complicates validity efforts as a straightforward endeavor. For instance, a user may have a belief that all web resources should have bullet-pointed recommendations in order to be useful in practice as opposed to blocks of text. Therefore, they may choose to adopt a non-evidence-based intervention because a web resource gave steps that were easy to follow. In this way, good intentions may nonetheless result in unintended outcomes.

Ing et al. (2021) provides an example of how the use of a measure designed to improve instruction in mathematics was adapted in two contexts, one that represents a productive use of data and one that does not. While the first context provides an example of how the measure was adapted for use in a productive way, the second context resulted in a teacher concluding that some of her students may not be capable of engaging in conceptually-oriented mathematics instead of procedural mathematics, citing in particular that 77% of students responded “yes” to the item *Was there only one right way to solve the problem(s) in class today?* This was a problematic interpretation in that it led the teacher to provide learning opportunities in conceptually-oriented math for only some students. The second context is an example of how data use can play out differently in practice than originally intended.

Guided by current understandings of best practice in validity arguments (see also Bell et al., 2012; Kane, 2009; Rino et al., 2021 for examples of applications), it is worthwhile to examine the actual interpretation of data resulting from measures within the context of web-based resources. In the case of ratings of usability of educational web resources, there is a need to investigate not just the intended use of the measure to assess the usability of web resources promoting interventions but also the actual use of the measure to assess the usability of web resources promoting interventions.

Usability

Usability plays a significant role in the adoption of an intervention, or in this case, web resource (Greenwood & Abbott, 2001; Riley-Tillman et al., 2005). Usability has been defined as “the extent to which a system, product, or service can be used by the specified users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use” (International Organization for Standardization, 2018). Although there are many possible inferences that can be made from web resources, this definition of usability lends itself to intervention uptake (Lyon et al., 2019) as usability taps the user’s perception of, in this case, the web resource’s ability to help them meet their goals in the classroom through implementation of a particular intervention. Intervention or lesson plan suggestions for teaching are also a few of the most common reasons that educators use web resources (Choi et al., 2018; Buren et al., 2021), even greater than social networking.

In relation to web resources, certain characteristics of usability are more applicable. Four characteristics resulted from a literature review by Mandracchia and Sims (2020).

> **“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." (Mandracchia & Sims, 2020, p. 8).

Usability offers one method for narrowing the research-to-practice gap by critically examining users' perceptions of research-based practices (Greenwood & Abbott, 2001), and is thus a crucial component of implementation science. Implementation science in turn is a field that is concerned with the adoption and promotion of evidence-based practices implemented with fidelity (Fixsen et al., 2009). Unfortunately, one of the exacerbators of the research-to-practice gap, and a foe of implementation science, is the "failure of research to produce many innovations that are usable in real classrooms" (Greenwood & Abbott, 2001, p. 281). This critique extends to web resources as well; teachers are less interested in using web resources that lack aesthetic appeal, feasible intervention recommendations, or accessibility (Buren et al., 2021). Thus, web resources that promote evidence-based practices in a way that is consumable for practitioners are likely integral to evidence-based practice adoption. In the next section, I will outline two existing measures of usability (i.e., the existing URP library and the SUS) and describe why a new measure is necessary to capture the full range of usability in relation to web resources.

Extant Tools for Usability Evaluation

Usability has been evaluated in different fields which have applicability to education-based research. Although the state of quality evaluation of web resources is somewhat limited, there are at least two tools for assessing usability which are relevant for this context.

System Usability Scale

One very widely cited measure is the System Usability Scale (SUS; Brooke, 1996), with the original paper having received over 12,713 citations as of February 21, 2022. The SUS is a ten-item measure of usability designed to create comparable evaluations of differing systems in an industrial context (Brooke, 1996). It is typically presented as a five-point Likert-type scale but can also be used on a seven-point scale (Brooke, 1996), with five positively worded items (odd numbered items) and five negatively worded items (even numbered items) that should be reverse coded (Brooke, 1996). The scale was designed to measure three aspects of usability: effectiveness (if users can complete the task and the outcome of that task), efficiency (the level of resources consumed to perform the task), and satisfaction (users' subjective reactions to using the systems).

The validation process for this measure included the use of 20 participants filling out a 50-item scale (Brooke, 1996). Participants were given one of two systems to evaluate, one that had been predetermined to be very user friendly and another that had been predetermined to be very non-user friendly. Items were retained based on strength of internal consistency, extremity of response, and ability to make a measure that had five positively worded and five negatively worded items.

One unique feature of the SUS is its wide use across disciplines. For instance, Arnab et al. (2015) used the SUS to evaluate the usability of their serious game (i.e., a game used for a purpose other than entertainment, such as a flight simulator for pilots), Ben-Zeev et al. (2014) used it to evaluate the usability of a smartphone intervention for schizophrenia, and Gupta et al. (2015) used it to examine the usability of TweetCred, a browser extension that assesses the credibility of Tweets.

Although adopted across several settings, some features of the SUS make it less optimal for the evaluation of educational web-based resources. For instance, factor analysis has demonstrated that it measures only one factor, *usability*, making it difficult to break down distinct usability components such as quality and feasibility (Bangor et al., 2008). Additionally, the scoring of the SUS is atypical, leading to confusion in scoring and interpretation (Bangor et al. 2008). Specifically, each of the ten items contribute a score of 0-4. Odd-numbered items contribute the scale position minus 1, while even-numbered items contribute 5 minus the scale position. The sum of scores is then multiplied by 2.5 to produce the total score. Furthermore, this measure was designed for use in an industrial context, as evidenced by items such as "I found the various functions in this system were well integrated," which is more difficult to relate to education. Although use has extended to educational contexts (e.g., Lyon et al., 2021), it is questionable whether there is sufficient evidence to support use in this manner. Relatedly, there has been little investigation into the actual use or inferences made as a result of the measurement data, indicating that use in practice may not be well understood in educational contexts (Ing et al., 2021). Therefore, other extant usability tools may provide additional insight into the measurement of usability.

Usage Rating Profile

The Usage Rating Profile (URP) library of measurement tools is designed to objectively evaluate perceptions of usability (Briesch et al., 2013). These instruments were constructed to help guide decision-making away from single-facet evaluations (such as acceptability or accessibility) and towards a more complete evaluation of the usability of the item being evaluated, whether that be an intervention or another measurement tool.

Resources available on the URP website include the URP Supporting Students' Behavioral Needs (URP-NEEDS), the URP-Assessment (URP-A), and the Children's Usage Rating Profile (CURP; UCONN, 2020). These tools contain a range of 21-29 items and have interpretation guidelines specifying how individual items map onto identified factors. One particularly relevant tool from this series is the User Rating Profile-Intervention Revised (URP-IR; UCONN, 2020), which gauges the usability of educational interventions across domains including *acceptability, understanding, feasibility, home school collaboration, system climate,* and *system support*.

The URP-IR was developed from a preliminary version of the measure known as the Usage Rating Profile-Intervention (URP-I; Briesch et al., 2013). To develop the URP-IR, Briesch et al. (2013) conducted an exploratory factor analysis on half of their sample (n = 503) and confirmatory factor analysis on the other half of their sample (n = 502) of in-service educators, which consisted largely of white females between the ages of 35-54 who held a master's degree or above and worked in a public school. Measure items were derived from existing intervention evaluation measures such as the Intervention Rating Profile 2.0 (Witts & Martens, 1983) as well as the original URP-I, with 75 total items included. Participants were contacted by phone, read a consent script as well as one of five vignettes describing an intervention, and then the initial items which participants responded to with the vignette in mind (Briesch et al., 2013).

For the exploratory factor analysis, eight factors were extracted. However, the researchers' decision rules indicated that items must have at least a 0.45 factor loading on their primary factor to be retained, and items in the seventh and eighth factor did not; thus a 34-item six-factor URP-IR was retained (Briesch et al., 2013). Five more items were deleted to improve internal consistencies within subscales, with final internal consistency values falling in acceptable ranges (α = 0.67-0.95; Briesch et al., 2013). The six factors were thus *acceptability* (9 items)*, understanding* (3 items)*, family-school collaboration* (3 items)*, feasibility* (6 items)*, system climate* (5 items), *and system support* (3 items). The confirmatory factor analysis also produced a six-factor model as the model of best fit, with acceptable fit statistics (χ^2 (74) = 383.63, χ^2/df = 5.18; RMSEA = .09, CFI = .96, SRMR = .05) as compared to a unidimensional model (χ^2 (62) = 2456.40, χ^2/df = 39.62, RMSEA = .30, CFI = .71, SRMR = .14; Briesch et al., 2013). Thus, the six-factor model was retained and constitutes the current URP-IR (Briesch et al., 2013). The development of the URP-IR thus far appears to collect evidence supporting the internal structure of the measure. A literature review did not reveal further inference checking; however, the URP-IR has been used in real educational contexts (e.g., Gilson et al., 2016; Payan et al., 2019; Chaffee et al., 2020).

The URP series of instruments has been used widely across educational research. For instance, Gilson et al. (2016) used the URP-IR to evaluate usability of their reading intervention, Payan et al. (2019) used the URP-A to evaluate usability of a curriculum-based measurement (CBM) assessment tool, and Chaffee et al. (2020) used both the URP-IR to examine teachers' perceptions of usability of a positive peer reporting intervention and the CURP to examine students' perceptions of the usability of that intervention. Although used across a number of educational contexts, some features of the existing body of URP measures make it less optimal for the evaluation of educational web resources. Most importantly, the measure's existing factors do not encompass important components of online resource utilization such as accessibility (Gunderson et al., 2006; Kelly et al., 2007; Mandracchia & Sims, 2020) and appearance (Jiang et al., 2016; Lawrence & Tavakol, 2006; Mandracchia & Sims, 2020; Tuch et al., 2010). Therefore, development of an evaluation tool that does account for these components would be a beneficial addition to the URP library.

Usage Rating Profile Web-Resource (URP-WR)

The usability of web resources was designed to be a key component in driving decision making about intervention selection and implementation (Beahm et al., in submission; Buren et al., 2021). Existing measures regarding quality of web resources exist (e.g., Shrock, SUS) but have key limitations in that inferences made to aid decision making have not been evaluated (Bangor et al., 2008; Kane, 2013; Ing et al., 2021; Mandracchia & Sims, 2020). Measures of usability exist but have not been extended to web resources.

Thus, the Usage Rating Profile-Web Resource (URP-WR) was created with the goal of filling this gap. The intended interpretation/use argument of the URP-WR is to aid decision making regarding whether a web resource is appropriate for use by educational professionals to inform their practice.

The quantification of scores is expected to allow for more objective, informed web resource evaluation. Additionally, the comparability of scores across different measures using a common scale allows for aggregate usability ratings to accompany web resource presentation. For example, an average item score per factor (i.e., accessibility, appearance, plausibility, and system support) and overall usability provide a way to compare different measures on the same scales. Another way to think of the inferences based on the URP-WR could be as a "Yelp" for web resource usability. Similar to "Yelp," users could have access to a rating accompanying a web resource they encounter. This could help them make inferences about use based on the factors evaluated by other users.

The URP-WR has already undergone a development process, and some work has already been conducted in order to evaluate its resulting data's structural validity and reliability (Mandracchia & Sims, 2020). This initial development was framed through an evaluation of evidence to support the measure in general, rather than the use of the measure in a way that attends to context and user specific interpretations. This study contributes to the ongoing development of this measure by attending to the key role users and context play. This study focuses specifically on two groups of users within the larger body of evidence consumers: doctoral students and pre-service teachers. The populations of doctoral students and pre-service teachers were chosen to represent researchers and practitioners in order to investigate how the two populations may interpret data differently. Researchers, specifically within school psychology, are explicitly and extensively trained in data-based decision making (Ysseldyke et al., 2006; Ysseldyke et al., 2008) while teachers and educational practitioners have varied levels of training (Labaree, 2018) as well as data use needs (Moss, 2013) and beliefs about their ability to use data (Datnow & Hubbard, 2015). It is important to the IUA of the URP-WR that both populations are understood.

Further, these populations of pre-service teachers and doctoral students, as opposed to in-service teachers and faculty, were chosen as these participants are likely to utilize web resources for decision making (Opfer et al., 2016), potentially due to their inexperience in the field (Sawyer & Meyers, 2018) or their level of comfort using online sources (Madden et al., 2005).

Previous URP-WR Development

Previous development of the URP-WR included initial item selection through consensus building and exploratory factor analysis (EFA) as well as initial measures of social validity (Mandracchia & Sims, 2020). The URP-WR is designed to aid researchers and practitioners in their decision of whether to use a web resource to inform their practice. The analyses in Mandracchia and Sims (2020) provided evidence regarding whether the URP-WR can be used as the developers intended, or more specifically, that the internal structure of the URP-WR is represented by four factors (Kline, 2016; Knetka et al., 2019). Evidence was also gathered from the Usage Rating Profile-Assessment (URP-A) on the face value (Connell et al., 2018) and perceptions of usability in their context (Lyon et al., 2021; Briesch et al., 2013).

Item Development and Reduction. Development of the URP-WR began with a literature review to determine what factors are critical to a usable web resource. The lack of peer-reviewed work on usability of web resources in education necessitated inclusion of non-peer-reviewed, publicly-available information in this review. Characteristics were drawn from the existing URP body of assessments, information provided by Kathy Shrock, and sources from fields such as management regarding the appearance and accessibility of web resources. This was an informal literature review, and the number of articles or resources consulted was not identified. Information gleaned from these sources resulted in common characteristics such as authorship, credibility, reliability, appearance (e.g., aesthetically pleasing), organization, accessibility, feasibility, technical components, and more. Commonalities between the characteristics resulted in the identification of four usability domains related to educational web resources: appearance, accessibility, credibility, and feasibility, which are further described in the usability section of this manuscript. Items were then developed relative to that informal literature review. The formatting of the items was modeled after the existing URP body of assessments. 112 initial items were developed based on the literature review, which were reviewed by the first author. After the review, 42 items were removed for redundancy and 70 items remained for the initial pool.

Of those 70 items, 15 were removed as those items failed to demonstrate 75% of raters (i.e., eight school psychology doctoral students and faculty from the University of

California, Riverside) sorting them into their hypothesized factors, indicating that the items may be confusing (Hennessy et al., 2016). The 55 remaining items were included in an EFA based on the results from 94 faculty, in-service educators, and undergraduate and doctoral students in the fields of education and psychology. 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).

In the initial EFA, items that had a pattern coefficient of less than 0.45 on their primary factor were removed, and items that had a pattern coefficient on a secondary factor above 0.30 were removed to avoid multidimensionality (Chafouleas et al., 2009). Based on these rules, 34 items were retained in the first factor analysis. A second EFA was completed to ensure all items met decision rules when ran again. Based on the decision rules, 31 items were retained from the second factor analysis.

For the final EFA with the retained items, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy reached a value of 0.79, indicating that the data were suitable to factor analysis. A four-factor structure emerged as guided by parallel analysis and a scree plot. Between four and six factors were suggested for extraction based on the scree plot (see Figure 3.2 of Mandracchia & Sims, 2020). Parallel analysis suggested extraction of four factors, with eigenvalues indicating that 55.2% of the variance in the data were explained using four factors. The four-factor structure explained 55.8% of the variance in the data, and 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 as the extraction of a fifth factor included only items that did not meet decision rules. Thus, although additional variance explained and same-to-better fit indices accompanied a five-factor model, four factors were extracted to eliminate redundancy (and because a four-factor model was suggested by the parallel analysis). All items then fell within the decision rules, and thus were retained in the final version of the URP-WR.

The results of the consensus building task and EFA provide initial evidence that the structure of the items seemed related to our conceptualization of how items related to the different factors. Comparison of possible models allowed initial confidence in the interpretation that usability is represented by four factors (i.e., plausibility, appearance, accessibility, system support), and that ratings on these factors can be used to make decisions about the usability of a web resource for a teacher or education professional's use in their practice.

Face Validity. Evidence regarding users' perceptions of the URP-WR was evaluated in order to understand how likely users are to interact with the URP-WR if provided. In other words, this evaluation was done to determine whether users perceived that they could make inferences from this measure in their practice; however, this was still conceptualized regarding intended interpretations rather than actual interpretations. This was evaluated using the URP-A (Chafouleas et al., 2012), which is a tool designed to measure the social validity of other measures.

The URP-A is a 28-item measure using a Likert-type scale with anchors ranging from 1 to 6. A selection of 23 items was used from the URP-A, as five (items 5, 7, 12, 15, and 27) did not apply to the IUA established for the URP-WR. Ratings among participants who responded were totaled and aggregated across usability domains. Although little guidance is provided for interpreting URP-A scores, higher scores are considered more favorable for the URP-A domains. Thus, the goal was an average overall score at or above 92, or an average rating of 4 out of 6 on Likert scale items, indicating that participants tended to positively agree with items. A secondary goal was 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 was an average item score of 4 per category.

76 total participants completed their ratings of the URP-WR using the URP-A. Participant composition reflected that of the larger sample used for the EFA (as 76 of the 94 elected to continue answering questions).

The overall average URP-A score across participants was 98.61, which was interpreted to indicate that participants perceived the URP-WR to be socially valid and acceptable to use in their setting. Six URP-A items fell under the category of *acceptability*; thus, the "best" score for this category would be 42 and the goal score was 28. The average score on this category was 29.89, and the average item score was 4.27. Three URP-A items fell into the category of *understanding*; thus the "best" score for this category would be 18 and the goal score was 12. The average score on this category was 13.37. The average item score was 4.46. Six UPR-A items fell under the category of *feasibility*; thus, the “best” score for this category would be 36 and the goal score was 24. The average score on this category was 27.31. The average item score was 4.55. Finally, four URP-A items fell under the category of *system climate*; thus the “best” score for this category would be 24 and 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 without additional system support). 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.

As acceptable results were obtained, it was concluded that participants (i.e., a sample of largely Hispanic and White Non-Hispanic female teachers, undergraduate, and doctoral students from the Southern California area primarily aged under 30) viewed the URP-WR as acceptable, understandable, feasible, and appropriate to the system climate, thereby providing initial evidence that the participants in this sample viewed the URP-WR as usable in their setting. The general characteristics of these users are reflected again in the current study, indicating evaluation of evidence for a similar user base.

Characteristics of the URP-WR

Thirty-one items were consistent with the four hypothesized factors: appearance (10 items), accessibility (5 items), plausibility (encompassing credibility and feasibility; 12 items), and system support (4 items). Plausibility contains 12 items that relate to the citations and believability of the information as well as the feasibility of the recommendations provided in the resource. Example items for this factor include “The resource cites its original sources" (from the credibility proposed factor) and "The resource contains all recommendations needed for implementation" (from the feasibility proposed factor). Appearance contains 10 items that relate to the overall design and appeal of the resource. An example item for this factor is "This resource is aesthetically pleasing." System support contains 4 items that relate to the support needed from administration or consultation in order to implement the recommendations provided in the resource. An example item for this factor is "I would need support from my administrator to implement recommendations made in this resource." Finally, accessibility contains 5 items that relate to the overall ease of accessing this resource on the internet. An example item for this factor is "It was easy to find this resource."

Participants respond to each of these items on a six-point Likert-type scale ranging from 1 (*Strongly Disagree*) to 6 (*Strongly Agree*). A high score on each of the scales was designed to indicate the following: (a) accessibility: the user perceived this resource to be easy to find and without roadblocks to accessibility, (b) appearance: the user perceived this resource to be aesthetically pleasing and thus easy to consume, (c) plausibility: the user perceived this resource as containing information from credible sources that could be easily understood and implemented practically, and (d) system support: the user would need more support from their system in order to implement the recommendations (the instructions on the URP-WR indicate that the user should consider reverse coding this item if they are looking for recommendations that can be implemented independently; Mandracchia & Sims, 2020).

Current Study

The next step in the validation process for inferences about usability based on the URP-WR is aimed at gathering evidence regarding actual use of the data. Going back to the example provided by Ing et al. (2021), actual use is important to understand so that adjustments can be made if actual use results in unproductive decisions regarding web resource usability (e.g., implementation of non-evidence-based interventions such as learning styles in a classroom because a web resource that was aesthetically appealing and easy to access recommended it). It is equally important to understand actual use to promote productive decisions regarding web resource usability (e.g., implementation of an evidence-based intervention in a classroom that a teacher found through a highly usable web resource). This study thus seeks to evaluate further evidence towards the URP-WR's IUA, focusing specifically on reported use/decision making regarding web resource uptake as well as prioritization of factors affecting decision making. The intervention itself is not key to the question, but rather the decisions made based on the URP-WR data.

Research Questions. This study asks five research questions.

1) How willing are users to implement an intervention based on given URP-WR scores?
2) Do pre-service teachers and doctoral students in educational psychology, special education, and school psychology make similar decisions about whether to implement a hypothetical intervention?

3) How do users prioritize different factors of usability in their decision whether to implement a hypothetical intervention?
4) Do pre-service teachers and doctoral students in educational psychology, special education, and school psychology prioritize similar factors in their decision making?
5) What might account for differences in user priorities of factors of usability?

Method

Participants

Two samples were targeted for participation in this study: pre-service or recently in-service (i.e., less than one year in the field) teachers at any grade level (i.e., did not distinguish between elementary, middle, and high school teachers or subject area), and doctoral students in school psychology, educational psychology, and special education. These samples were chosen to represent two different types of potential URP-WR users. As discussed earlier, researchers and teachers may use data differently due to differences in training as well as underlying beliefs. Further, populations of people in training were chosen as they are more likely to utilize web resources than those with much experience in the field or who may not be as comfortable using online sources (Hargittai et al., 2019; Hunsaker & Hargittai, 2018). Although pre-service researchers are indeed distinct from in-service researchers, they are in training to become researchers and may be considered to be more closely related to researchers than teachers. All participants were over 18 years old and fluent in English. The sample from which participants were recruited reflects a convenience sample.

Participants were recruited through emails to the first author's contacts and via word of mouth or email from instructors or teaching assistants. 20 school psychology doctoral students from the University of California, Riverside (UCR) were emailed directly to recruit study participation. Additional UCR doctoral students in special education and educational psychology were recruited through two UCR faculty, who emailed recruitment information to their doctoral students after being contacted. Doctoral students from other universities were recruited through emails to seven program directors of school psychology programs, who were in my contacts as they had agreed to send recruitment information to their students for a previous study. The programs solicited were the University of Denver, the University of Montana, the University of Colorado at Denver, the University of Kentucky, the University of California at Santa Barbara, the University of California at Berkeley, and the University of Oregon. Of those seven program directors, six responded agreeing to send the recruitment information to their doctoral students. Special education doctoral students were also solicited from the University of Virginia through an email to a program alumna. All doctoral programs that were solicited have a research focus, while the UCR teacher education program has a very strong practitioner (i.e., teacher) focus.

Pre-service teachers were recruited via email to two instructors in the teacher education program at UCR, who sent emails and made announcements in Canvas with recruitment information. An additional 56 pre-service teachers from UCR were recruited via email. Finally, pre-service teachers and doctoral students were recruited from UCR through a department-wide email from the graduate education department coordinator.

Participants reported demographic information consisting of age, gender identity, race, ethnicity, and role (e.g., pre-service teacher or doctoral student). Efforts were undertaken to exhaust personal networks (see above), and the use of this convenience sample may not generalize beyond Southern and Northern California pre-service teachers and doctoral students.

A power analysis to estimate the number of participants for this study was not conducted. The primary rationale for not conducting a power analysis is that there is no similar previous research upon which to base parameters for such an analysis. In addition, given the resources available, it was not feasible to more broadly recruit participants or to highly incentive participation. Instead, a minimum of 30 participants per group (33 doctoral students, 37 pre-service teachers) were recruited.

A total of 138 people provided consent to participate in this study. Four participants provided consent but dropped out completing no demographic information. An additional 39 participants were excluded from participation as they did not identify as either a pre-service or recently in-service (i.e., less than one year in the field) teacher or a doctoral student in school psychology, educational psychology, or special education. An additional 15 participants dropped out after the demographic portion and did not complete the study. Eight participants (two doctoral students, six pre-service teachers) were excluded from data analysis due to irregular response patterns (e.g., moving the number "1" in the ranking question one place each time, their rankings did not match their qualitative response for reasoning behind rankings), and two participants (pre-service teachers) were excluded as their responses were exact duplicates including number of practicum hours. Of the 70 participants whose data were used in analysis, 37 of those participants were pre-service teachers recruited from the UCR teacher education program and 33 were doctoral students recruited from UCR, the University of California, Berkeley, the University of California, Santa Barbara, and the University of Oregon school psychology, educational psychology, and special education doctoral programs.

The majority of participants were female (90%), White (51%), and non-Hispanic (51%; see Table 1). There was no significant difference between groups in terms of gender, χ^2 (1, N = 70) = 0.91, p = 0.33. However, there was a significant difference between conditions in race, χ^2 (6, N = 70) = 15.63, p = 0.02 as well at ethnicity, χ^2 (1, N = 70) = 7.01, p = 0.01. Demographic categories for race and ethnicity reflecting the U.S. Census categories were used in this study. It appears that pre-service teachers and doctoral students interpreted the race demographic item differently. The US Census and many research project demographic data collection procedures consider race and ethnicity to be separate categories, meaning those of Hispanic origin would select White (or another race that applies such as American Indian or Alaska Native) for their race and Hispanic for their ethnicity. Doctoral students selected White as their race and Hispanic as their ethnicity, while pre-service teachers appear to have selected Other and filled in Mexican/Mexican American or Latino/a/x (n = 10) for race and selected Hispanic as their ethnicity. Therefore, race/ethnicity differences among this sample should be interpreted with caution. It is also important to note that categorizing race in any way, but especially this distinction, is a limitation recognized by the U.S. Commission on Civil Rights (USCCR, 2002).

About one third of total participants responded “Other” for race/ethnicity, which included text responses such as Middle Eastern, Multiracial (no further specification), Latino/a/x, and Mexican/Mexican American. Participants had a mean age of 26.33 (SD = 5.41; see Table 2). The doctoral student population was significantly older than the pre-service teacher population (*t* = 2.74, *p* = 0.008).

Measures

URP-WR

The URP-WR (Mandracchia & Sims, 2020) was developed to measure the usability of educational web resources. The URP-WR is currently in its initial validation stages and was derived from the URP-IR (Chafouleas et al., 2011) as well as currently available evaluations of web resources (i.e., Schrock, 2020; Lydia M. Olson, 2018) with the aim to aid educational researchers and practitioners in making decisions about the usability of web resources for their setting relative to intervention/lesson plan uptake for classroom use. Notably, in this study, participants did not actually fill out the URP-WR, but rather reported their Perceptions of Use based on scenarios containing hypothetical URP-WR data.

Perceptions of Use

To measure inferences that participants make based on scores on the URP-WR, participants were asked if they would be willing to implement a new intervention designed to improve their teaching, responding dichotomously (i.e., yes or no). Doctoral students were asked if they would use this in their teaching/TA’ing practice in order to most closely mirror the decision made by pre-service teachers. The instructions appeared as follows: “Imagine that you had access to the usability average ratings of a website providing information on a new type of intervention designed to improve teaching (graduate level teaching/TA’ing) from 1000 teachers (doctoral students) nationwide. Take between 2 and 5 minutes to review the data provided below, then answer the following questions.”

Participants were then asked to rank the factors that they prioritized from 1-5. For the Perceptions of Use, the plausibility factor is split back into original hypothesized factors of credibility and feasibility as these represent distinct constructs, as well as that the factor analysis that suggested inclusion in one factor was based on a small sample size (*n* = 96) based upon one resource which may have been viewed as both highly credible and feasible. Therefore, participants were asked to rank the factors of feasibility, credibility, accessibility, appearance, and system support from 1-5 with 1 being most influential and 5 being least influential to their decision making.

Participants were provided the Perceptions of Use five times accompanying five presentations of data (i.e., Scenarios, see next section). The order of Scenario presentation was counterbalanced across participants using Qualtrics’ “counterbalance” feature to mitigate possible order effects.

After the last Scenario was presented to them, participants were asked to respond narratively to the question “Thinking back to all five scenarios, why did you rank the factors in the way that you did? Please give a brief (1-2 sentences) explanation of your general reasoning.” See Appendix B for the Perceptions of Use and directions presented to participants.

Scenarios. The Perceptions of Use items were accompanied by five scenarios. A description of each factor (i.e., feasibility, credibility, appearance, accessibility, system support) and the average item score for each factor resulting from a hypothetical body of 1000 teachers (or doctoral student to match the participant's role) was provided for a hypothetical web resource relating to a hypothetical new intervention to improve teaching (or graduate level teaching/TA'ing). These web resources all pertain to this same intervention but the scores for the intervention varied. The intervention chosen was not considered to be crucial to the study, but rather the participant's decision to implement the intervention based on differing levels of URP-WR factors. The average item score was presented out of 6, as the URP-WR items are presented on a 1-6 Likert-type scale.

The five scenarios are designed to represent five possible website types that participants may encounter: (a) high on all five factors, (b) low on all five factors, (c) medium on all five factors, (d) high on credibility, low on appearance, accessibility, feasibility, and system support (e) low on credibility, high on appearance, accessibility, feasibility, and system support. The distinction between credibility and the other factors is designed to mimic the dichotomy between (a) websites that provide evidence-based but difficult-to-implement recommendations on an unattractive user interface and are difficult to find through, for example, a Google or TPT search versus (b) those that provide non-evidence-based recommendations that are very easy to implement on an aesthetically pleasing site that has high search engine optimization and may be one of the first or second results from, for example, a Google or TPT search. The five specific scenarios to be used in this study are outlined in Appendix C.

Procedures

This study was reviewed and approved by UCR's Institutional Review Board before initiating participant recruitment.

A Qualtrics link and QR code were embedded in the recruitment email to the first author's contacts and was provided by the teacher/teaching assistant for participants recruited through coursework. When an interested party clicked on the link or scanned the QR code, they were directed to an information sheet and were only allowed to complete the study upon reviewing the sheet and indicating consent to participate. Participants were informed that they were expected to answer 5 demographic items, 3 items relating to their teaching experience, and 11 items relating to their decision-making in a hypothetical scenario regarding how to use an educational website in practice; the term "website" was used as this was considered to be more likely to be familiar to participants than "web resource." Participants were informed that the study was expected to take approximately 15-25 minutes to complete, and that they were allowed to take breaks as needed. They were informed of minimal possible risk associated with participation in the study, including eye strain due to looking at a computer screen and time loss for study completion. They were informed that entry into a raffle for one of five $20 gift cards may be obtained upon completion of the study, or by requesting it through email to the principal investigator.

After providing informed consent, participants were asked to provide their demographic information including age (provided numerically by the participant), gender identity (Cisgender Man, Cisgender Woman, Transgender Man, Transgender Woman, Non-Binary/Third Gender, or Other [narratively completed]), race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Multiracial, or Other [narratively completed]), ethnicity (Hispanic or non-Hispanic), and role (Doctoral Student, Pre-Service or In-Service [less than one year in the field] Teacher).

Depending on their reported role, participants were asked about their teaching experience or graduate level teaching/TA'ing experience. Doctoral students were asked how many classes they have TA'd for without leading a discussion or lab (on a scale from 0-20), how many classes they have TA'd for a led a discussion/lab (on a scale from 0-20), how many classes they have been an instructor of record for (on a scale from 0-20), and which populations they have TA'd/been instructor of record for (undergraduate, graduate, other [narratively completed by participant]). Pre-service/in-service teachers were asked if they had direct classroom experience in a teaching capacity including supervised practicum hours (yes/no); if yes, how many quarters of practicum hours had they completed (provided numerically by the participant); if they have completed edTPA required independently implemented lessons (yes/no); and how they would rate their level of independence in the classroom from 1-5 (Likert-type scale).

Participants were then presented with the first scenario and the Perceptions of Use items outlined in the Measures section. They were instructed to imagine that they had access to the average item ratings of 1000 teachers (or doctoral students to match their own characteristic) for a website providing information on a new intervention designed to improve their teaching (or graduate level teaching/TA'ing) for five categories of usability, to take between 2-5 minutes reviewing the data and then answer the questions below (i.e., Perceptions of Use). Each category was defined next to its rating (see Scenarios). The rating was provided as an average of items in that category on a scale from 1-6 for each factor. The four subsequent scenarios were presented in the same manner, with instructions to now imagine that this new set of data (i.e., the new Scenario) accompanied the website instead. They completed the last item after rating all the scenarios.

After completing the study, participants were redirected to a separate link to submit their email address to be entered into a raffle for one of five $20 Amazon gift cards. The $100 needed to fund this incentive was provided by the Trainers of School Psychologists (TSP) Graduate Student Scholarship. This separate link was created to allow participants to retain the anonymity of their responses. Participants were informed they may also send an email requesting entry as outlined in the information sheet, per IRB guidelines. Upon completion of data collection, five email addresses were selected at random and sent a $20 Amazon gift card.

Data Analysis

Research Question 1

We examined descriptive statistics pertaining to willingness to use the web resource to respond to research question one. We specifically examined the frequency and percentage of participants who endorsed being willing to use the website to help them implement the intervention in their classroom/TAship for each Scenario. We also provided a table representing these statistics. These statistics were used to understand how many overall participants report willingness to use the web resource in different Scenarios. This information is important in order to get a full picture of participants' willingness to use web resources based on usability data.

We also provided a bar graph for visual representation of the data. This bar graph is provided with bars representing the percentage of total participants in each scenario who indicated that they would use the web resource to help them implement the intervention in practice. This is provided in order to allow for easy visual analysis and comparison of differences in web resource promotion in different Scenarios.

It was hypothesized that participants would be willing to use the web resource to help them implement the intervention for Scenario A but not for Scenario B. The distinction between Scenarios C, D, and E was hypothesized to be more variable, with potential differences in willingness to use interventions depending on role as well as Scenario (see Research Question 2).

Research Question 2

We used two-sample *t*-tests to determine whether there was a relationship between participant role (i.e., pre-service teacher or doctoral student) and whether they indicated they would use the web resource to help them implement the intervention. Prior to conducting *t*-tests, the data were tested for assumptions of normality using Shapiro tests (shapiro.test function in R) and equality of variance using Levene's tests (leveneTest function in R). The Shapiro and Levene's test were statistically significant, meaning that the assumptions of normality and homogeneity of variance were violated. Therefore, these results should be interpreted with caution. Outliers were checked using the 1.5 x

IQR rule, and none were found. Finally, the assumption of independence of observations was met as each subject only belongs to one group, but the assumption of random sampling was not met.

A table representing the mean, *t* statistic, and p-value for a participant's endorsement of being willing to use the web resource to help them implement the intervention in their practice by (a) pre-service teachers and (b) doctoral students was provided. It was hypothesized that there would be no significant difference between the mean level of endorsement of intervention usage by pre-service teachers and doctoral students for the first two scenarios, such that:

H_0: there are no differences between pre-service teacher and doctoral student responses where dimension variability is low. $\mu_1 = \mu_2$

H_1: There are differences between pre-service teacher and doctoral student responses where dimension variability is low. $\mu_1 \neq \mu_2$

In the first two scenarios, the Scenario-provided data either unanimously support or do not support usability. Therefore, it was hypothesized that users would indicate willingness to use the website to help them implement the intervention when the Scenario-provided data supports it, but not when it does not.

Scenarios C, D, and E depict more variability across URP-WR dimensions. It is plausible that pre-service teachers would be more likely to endorse trying the intervention compared to doctoral students in Scenarios C, D, and E, as doctoral students are more likely to have been recently trained to go into depth when researching a new practice due to their training in data-based decision making and regular interactions with research

(Ysseldyke, 2006; Ysseldyke, 2008). It is also possible that pre-service teachers would be more likely to endorse intervention usage in Scenario D and less likely in Scenario E. Previous research has demonstrated that teachers report making use decisions based on accessibility and appearance, as well as frustration with the roadblocks that they experience with resources that lack these features (Buren et al., 2021).

Since the exact directions of expected effect are unclear, two-tailed *t*-tests were conducted. In other words, it was hypothesized that the null hypothesis will be rejected and that the means will not be equal (i.e., $\mu_1 \neq \mu_2$), but there is no specific direction hypothesized (e.g., $\mu_1 > \mu_2$) where:

H_0: there are no differences between pre-service teacher and doctoral student responses where dimension variability is high. $\mu_1 = \mu_2$

H_1: There are differences between pre-service teacher and doctoral student responses where dimension variability is high. $\mu_1 \neq \mu_2$

In addition to *t*-tests, another bar graph was utilized to aid interpretation of results, this time with two bars representing the percentage of (a) pre-service teachers and (b) doctoral students who indicated that they would use the web resource to help them implement the intervention for each scenario (see Figure 2 with hypothetical data). This allowed for visual representation of the data in addition to statistical tests.

Research Question 3

We provided a table representing frequency and percentage of participant factor rankings. This allowed for comparison of overall participant factor rankings within the table. Further differences in factor rankings are addressed through visual analyses of figures.

We also provided bar graphs to represent the participant rankings of how each factor impacted their decision making: five bar graphs by factor, representing the percentage of participants who ranked the factor first, second, third, fourth, and fifth in each Scenario. In addition, we provided bar graphs to represent the participant rankings of the first and last ranked factor: one bar graph representing the percentage of participants who ranked each factor first in each Scenario, another bar graph representing the percentage of participants who ranked each factor last in each Scenario. These bar graphs are provided to allow for additional insight into differences between Scenarios through visual comparison.

Overall, it was hypothesized that participants would prioritize feasibility and credibility over the other factors of the URP-WR. However, data use research suggests that people are highly influenced by their beliefs and previous knowledge when making decisions (Coburn et al. 2009a; Coburn & Turner, 2011). Therefore, factors that influence their decision may also be variable by participant background and knowledge.

Research Question 4

We provided two bar graphs representing the frequency and percentage of (a) pre-service teacher and (b) doctoral student rankings. We provided additional bar graphs by factor, this time with two bars for each Scenario, separated by participant role. Similar to those broken down by factor, we provided additional bar graphs for separate populations of pre-service teachers and doctoral students, meaning there are four additional bar graphs (e.g., pre-service teachers who ranked each factor first, doctoral students who ranked each factor first, pre-service teachers who ranked each factor last, doctoral students who ranked each factor last). It is important to have these comparisons in order to understand our populations of interest at opposing sides of the research to practice gap, and to understand how to best fit their data use needs.

It was hypothesized that while participants will prioritize feasibility and credibility over the other factors (see Research Question 3), doctoral students will prioritize credibility over feasibility, but pre-service teachers will prioritize feasibility over credibility. This is based on training in data-based decision making that occurs for researchers or research track professionals in general (Ysseldyke et al., 2006; Ysseldyke et al., 2008), as well as the fact that teachers have varied levels of training (Labaree, 2018) and data use needs (Moss, 2013) with feasibility being a priority for the sake of time and resources which can be very limited for teachers (Teig et al., 2018; Vannest & Hagan-Burke, 2010).

It is important to understand the prioritization of factors so that, if necessary, weights can be given to factors when calculating an overall usability score in URP-WR implementation. For example, teachers may use a different form of the URP-WR than researchers in order to more heavily weight credibility and move away from poor decisions made based on data from this measure. Similarly, researchers may prioritize credibility so much that they overlook the need for the intervention to be feasible in their setting and may end up making poor intervention decisions by taking on more than they are capable of doing in their setting. Both, neither, or other results may arise from this Research Question.

Research Question 5

Narratively completed responses to the last item asking participants why they ranked the factors the way they did were analyzed first through development and organization of codes based on words and phrases into categories based on established patterns (Creswell & Clark, 2018). Themes were then derived from these patterns (Krueger, 2014). We provided the number of participants whose responses fell into each theme.

Results

Research Question 1

Research question 1 asked how willing users are to implement an intervention based on URP-WR scores. It was hypothesized that participants would willing to use the web resource to help them implement the intervention for Scenario A but not for Scenario B. On the Perceptions of Use item 1, sixty-six (94%) participants indicated that they were willing to use the web resource to help them implement the intervention in their classroom/TAship for Scenario A (high scores on all factors). Fifteen (21%) participants did so for Scenario B (low scores on all factors). Thus, the results suggest failure to reject this hypothesis.

The distinction between Scenarios C, D, and E was hypothesized to be more variable, with potential differences in willingness to use interventions depending on role as well as Scenario. Twenty-eight (40%) participants did so for Scenario C (medium scores on all factors). Twenty-six (31%) participants did so for Scenario D (high score on credibility, low scores on the rest). Finally, twenty-one (30%) participants did so for Scenario E (low score on credibility, high scores on the rest). See Table 3 for a table representing this data, and Figure 1 for a bar graph representing this data. The results for these Scenarios suggest failure to reject this hypothesis as variability is evident, but lingering questions remain about differences in intervention endorsement by role.

Research Question 2

To address differences by participant role, research question two asked whether there is a relationship between participant role (i.e., pre-service teacher or doctoral student) and if they would use the web resource to help them implement the intervention.

In the first two scenarios, the Scenario-provided data either unanimously support or do not support usability. Therefore, it was hypothesized that users would indicate willingness to use the web resource to help them implement the intervention when the Scenario-provided data supports it, but not when it does not. The null hypothesis was not rejected for Scenario A as there was no statistically significant difference between pre-service teacher and doctoral student endorsement of using the web resource to implement the intervention for Scenario A (high on all factors; $t(36) = 1.78$, $p = 0.083$; see Table 4) which was predicted; however, the null hypothesis was rejected for Scenario B as there was a statistically significant difference between pre-service teacher and doctoral student responses for Scenario B (low on all factors; $t(36) = -3.22$, $p = 0.002$) which was not predicted. In this Scenario, more pre-service teachers endorsed usage of the web resource to help them implement the intervention in comparison to doctoral students.

Scenarios C, D, and E depict more variability across URP-WR dimensions. It is plausible that pre-service teachers would be more likely to endorse trying the intervention compared to doctoral students in Scenarios C, D, and E, as doctoral students are more likely to have been recently trained to go into depth when researching a new practice due to their training in data-based decision making and regular interactions with research (Ysseldyke, 2006; Ysseldyke, 2008). Therefore, it was hypothesized that there would be significant differences in endorsement of using the web resource to try the intervention between pre-service teachers and doctoral students.

The results suggest failure to reject the null hypothesis for Scenario C (medium on all factors; $t(36) = -1.07$, $p = 0.288$) and Scenario D (high on all factors but credibility, $t(36) = -1.12$, $p = 0.268$) which was not predicted; however, the null hypothesis was rejected for Scenario E (low on credibility, high on the rest of the factors; $t(36) = -2.10$, $p = 0.040$), which was predicted. Similar to Scenario B, more pre-service teachers endorsed using the web resource to help them implement the intervention. However, that was the predicted outcome for Scenario E but not for Scenario B.

In addition to *t*-tests, another bar graph (see Figure 2) was utilized to aid interpretation of results, this time with two bars representing the percentage of (a) pre-service teachers and (b) doctoral students who indicated that they would use the web resource to help them implement the intervention for each scenario.

With research questions 1 and 2 addressed regarding intervention uptake using URP-WR data, additional questions remain regarding how users made those decisions. Specifically, how did users prioritize URP-WR factors in their decision making?

Research Question 3

Research Question 3 addresses prioritization of factors in decision making of all participants. It was hypothesized that participants would prioritize feasibility and credibility over the other factors. Visual analysis of graphs (Figures 3-9) representing this data indicate that participants prioritized credibility first most often (58% of participants ranked it first in Scenario A, 60% in Scenario B, 55% in Scenario C, 42% in Scenario D, 78% in Scenario E; see Table 5). Across all scenarios, credibility was the most important factor in decision-making regarding whether to use a web resource to implement an intervention. In general, the trend among Scenarios stayed the same for rankings of factors. However, Scenario D had fewer participants rank credibility first (42%) compared to Scenarios A, B, & C (58%, 60%, 55%) while Scenario E had more participants rank it first (78%). Scenarios D and E manipulated credibility the most strongly, which makes these findings interesting. It appears that when credibility is known to be strong, it may affect participants' decision-making less than when it is known to be weak.

Accessibility was largely prioritized higher (average 21% prioritized first, 28% prioritized second across Scenarios) than feasibility (average 11% prioritized first, 34% prioritized second across Scenarios), which was not predicted. Appearance (average 46% prioritized last across Scenarios) and system support (average 38% prioritized last across Scenarios) were least prioritized, which aligns with the hypothesis.

The general consensus indicates that credibility is the most highly prioritized factor, leading to a subsequent question of whether there are differences in prioritization between pre-service teacher and doctoral student responses, thereby potentially affecting how different target populations use the URP-WR data.

Research Question 4

Since data use research suggests that people are highly influenced by their beliefs and previous knowledge when making decisions (Coburn et al. 2009a; Coburn & Turner, 2011), research question 4 addresses differences by role (i.e., pre-service teacher vs. doctoral student) in prioritization of URP-WR factors.

It was hypothesized that doctoral students would prioritize credibility over feasibility, but pre-service teachers would prioritize feasibility over credibility. Visual analysis of graphs representing these data (see Figures 10-16) suggest that this hypothesis was not supported. Pre-service teachers and doctoral students both prioritized credibility most often (see Tables 6 & 7).

Pre-service teachers prioritized accessibility higher than feasibility more often, while doctoral students prioritized feasibility over accessibility more often. Both populations ranked appearance and system support as the least influential factors, with pre-service teachers more often prioritizing appearance over system support and doctoral students more often prioritizing system support over appearance.

Research Question 5

Out of the 70 total participants, 66 provided narrative responses. Three pre-service teacher responses that were not interpretable (e.g., “I went for what people look for”) were not analyzed, leaving sixty-three responses.

49 (77%) participants indicated in their responses that credibility was the most influential factor in their decision making, which corresponds with the numeric ranking results. 10 out of those 49 (20%) mentioned credibility as the sole factor that affected

their decision making. Four of those ten (40%) were pre-service teachers, and six of the ten (60%) were doctoral students.

An example of a doctoral student response that focuses solely on credibility is as follows, “I focused most on credibility as a deciding factor because I would prefer not to use an intervention that does not have an evidence base. I am also confident in my ability to navigate most problems.” An example of a pre-service teacher response that focuses solely on credibility is as follows, “Credibility is really important and should be the bases for everything in teaching. It would be pointless to not have credible and reliable sources in research.” This finding largely supports credibility being the most influential factor, as found in factor rankings.

Few participants mentioned other factors in their narrative responses as being the most important in their decision making. However, 10.60% of participants indicated that appearance was the least influential factor, or that they did not consider it at all. This falls in line with the numeric factor rankings.

Few participants showed deeper levels of critical thinking than naming most or least important factors, making true analysis for research question 5 difficult. However, two participants (both doctoral students) indicated a general desire to achieve a balance of all factors but nevertheless acknowledged that one (credibility) played more heavily onto their decision making than others. An additional two participants (both doctoral students) indicated that their factor rankings depended on the scenario. For example, in the case of low rankings (a generally lowly usable resource), one participant prioritized credibility but in the case of high rankings (a generally more usable resource), they prioritized accessibility. The other participant prioritized credibility but was willing to overlook high credibility if the other factor rankings were unusable.

Discussion

Users Make Decisions Based on Aggregate URP-WR Data

Based on the results of this study, potential users of the URP-WR (i.e., pre-service teachers and doctoral students) indicated they would use a web resource to help them implement an intervention to improve their teaching when aggregate URP-WR results unanimously support doing so, and would not use the web resource when aggregate URP-WR results unanimously discourage from doing so. This demonstrates that in extreme cases, the URP-WR can be helpful to decision making.

When aggregate ratings were mixed, results indicate that participants responded differently. A web resource that was medium on all factors was more likely to be used than one that is uncredible but has high ratings on the rest of the factors as well as one that is credible but low on the rest of the factors. This indicates that users would rather use a resource that is mediocre across factors than one that has severe flaws, which is promising.

These results most importantly demonstrate that users take aggregate ratings from other users who match their characteristics into account when making intervention decisions. This aligns with data use research indicating that others' perceptions can influence decision making (see Fogel & Zachariah, 2017; Luca, 2016). Therefore, providing aggregate ratings in "Yelp" style in conjunction with What Works

Clearinghouse, Teachers Pay Teachers, or another website could be potentially informative uses of URP-WR data.

Users Prioritize Credibility in Decision Making

Potential users of the URP-WR prioritized credibility most highly across scenarios in numeric factor rankings as well as narrative responses to the question, “why did you rank the factors the way that you did?” This indicates that when users are forced to consider credibility as a salient factor, they prioritize it very highly in their decision making. This was true across both doctoral students and pre-service teachers, indicating that both are likely being trained to value credibility. However, nearly a third of participants indicated that they would still use a web resource that is not credible to inform intervention implementation. Although this is ultimately the user’s decision, it may make sense to consider weighting credibility scores more heavily while using the URP-WR in this context.

Alternatively, it may make sense to emphasize some factor rankings over others. Although an initial literature review gleaned four salient factors (Mandracchia & Sims, 2020), it seems that current data indicate a very distinct preference for some factors over others in regard to decision-making practices. For example, if only three ratings can be displayed, it may be best to use credibility, accessibility, and feasibility over appearance and system support. Or it may be better to include system support and appearance items as supplemental which would allow for a shorter version of the primary URP-WR. Most drastically, it may even be beneficial to only display credibility ratings. This would no longer be a measure of usability, and thus would need to reflect that change. However, given the possibility of decision overload (Buchanan & Knock, 2001) with too much information, coupled with the fact that users overwhelmingly prioritized credibility, this could be a useful extension of the URP-WR. These use practices would need to be empirically tested to determine their validity in each setting

If usability continues to be the direction of the URP-WR, it may be beneficial to include case studies with the URP-WR to provide examples of different/deeper ways of thinking about the data, as only four participants endorsed a deeper level of thinking regarding factor prioritization when interpreting URP-WR results. This would be more difficult for the aggregate rating scenario, but it could be implemented in research feasibly. Given the overall results pointing to the usefulness of aggregate data, this direction appears to be non-ideal. Nonetheless, understanding deeper levels of thinking about data-based decision making is an interesting future direction for research.

Using the URP-WR to Narrow the Research to Practice Gap

This study investigated two distinct populations of potential URP-WR users. In general, pre-service teachers and doctoral students both used the URP-WR and prioritized factors (i.e., prioritized credibility most highly) similarly. However, there were also some sensible differences. First, pre-service teachers were significantly more willing than doctoral students to use the web resource when aggregate ratings were low on all factors. Second, pre-service teachers were significantly more willing than doctoral students to use the web resource when aggregate ratings are low on credibility and high on the rest of the factors. These results demonstrate the importance of facilitating communication between researchers and practitioners, as their interests are not always aligned.

Although the URP-WR is by no means sufficient to ensure that researchers and practitioners are exclusively using credible web resources and interventions, it can help by making sure that one way to evaluate credibility is accessible and salient. This may help close the "research-to-practice gap" (Carnine, 1997) in that practitioners will have greater access to information about the credibility of a web resource they plan to use. The URP-WR can also be a way to initiate conversations between researchers and practitioners. Practitioners could be provided aggregate ratings from researchers, especially regarding credibility. Researchers, on the other hand, could be provided aggregate ratings from practitioners, especially regarding feasibility or accessibility. This is an interesting direction for future study.

Limitations

This study has limitations that should be noted when interpreting the results. The use of a convenience sample and the sample size indicate that results can primarily be generalized to populations that closely resemble the population studied (i.e., pre-service teachers and doctoral students in the Southern California). Additionally, the limitation of categorizing race means that the true diversity of the population may not have been accurately captured, as mentioned in the Participants section of this manuscript. Further, the inferential statistics drawn from this analysis should be interpreted with caution as the participants were not randomly sampled from the population, and the assumptions of normality and homogeneity of variance were violated. Future research should replicate studies investigating use decisions with a larger, more diverse, random sample using more fluid racial demographic categories.

Although there was a strong rationale for focusing on pre-service teachers and doctoral students, it must be acknowledged that this was a sample of convenience. A wider population of potential users who may benefit from the URP-WR was not investigated in this study. For example, it is not known the extent to which doctoral students in education represent researchers more broadly. In addition, there is also great variation in pre-service teachers which could not be explored within the scope of this study. Thus, there may be a difference in how experienced teachers and researchers make use of this data (Eggleston, 2018) that require greater attention on a more representative scale. These populations should be explored in future studies in order to determine whether the URP-WR can be validly used within these populations.

There is also a procedural limitation in that pre-service teacher were asked to rate their level of classroom independence while doctoral students were not. This means that this metric could not be compared across populations. This was done so that pre-service teachers and doctoral students would be answering the same number of demographic questions, but it was an oversight that missed an interesting comparison. Future research should consider classroom independence across populations.

Additionally, as this study measures reported use decisions rather than in vivo use decisions, demand characteristics (when a participant responded to items in the way they feel they are “expected” to respond or based on perceived social desirability; see Orne, 1962; 1996; McCambridge et al., 2012), may bias responses from what would be seen in an authentic setting. Future research should task participants with making proposed decisions in order to determine their use in practice. This would also take other intervention adoption/implementation factors (e.g., time constraints) into account, more closely mirroring a real world scenario.

Importantly, this study investigates one type of use: an intervention designed to improve teaching practices. This was by design as specificity in use cases makes stronger validity arguments. However, this also means that future research needs to be conducted with web resources promoting other types of interventions (e.g., to be used with students individually, to be used in a tiered system) as well as for other education practices that do not involve interventions. Although some extension of the results is possible, higher-stakes decision-making practices may (and should) be accompanied with more caution from users. Indeed, one participant in their narrative response wrote that "any practice that will improve my skills should be tried," and it is unclear whether that would extend to student skills, for example.

Finally, this study investigated only factors derived from one measure of web resource usability. Other possible latent factors affecting web resource usability, for example coverage (e.g., are these topics successfully addressed, with clearly presented arguments and adequate support to substantiate them? Is the target audience identified and appropriate for your needs?), should be explored further. The separation of the initially proposed factors of feasibility and credibility from their combined factor, plausibility, may also affect decision making and should be further explored. Therefore, although this study provides initial information regarding use decisions in a particular context (i.e., intervention uptake), results should be interpreted only in that context and further research is needed to warrant additional inferences.

Conclusion

Overall, this study sought to evaluate evidence of decision making resulting from the URP-WR. Specifically, this involved the evaluation of evidence regarding perceptions of the actual use of the measure in decision making. Recently, the importance of focusing on both intended and actual interpretation of instrument-derived data, more specifically those inferences made by the user, has become a topic of extensive focus (e.g., Cizek, 2016; Haertel, 2013; Ing et al., 2021; Shepard, 2016). This study sought to contribute to this body of literature by focusing on the perceptions of the actual use of a measure aimed at supporting data-based decision making in the use of educational web resources.

This study suggested that users may make informed decisions based on aggregate data, and that they prioritize credibility most highly out of factors gleaned regarding web resource usability. This implies that future directions of the URP-WR should take aggregate use into play, as well as potential need for limiting factors to meet the needs of its users. This study also distinguishes between two diverse populations of users: pre-service teachers and doctoral students. Although both populations valued credibility, pre-service teachers were more likely to endorse willingness to implement an intervention in its absence.

Although most scholars agree on the importance of utilizing evidence-based practices, ways to promote them are less understood. One way to advocate for the implementation of evidence-based practices is to evaluate the usability of educational resources available on the internet, one of the most-used sources of in-service practices for educators. Evaluation and data allow users to select resources that best suit their needs and make appropriate use decisions, developers to solicit feedback to improve their resources, and the field to advance in implementation of evidence-based practice. The URP-WR allows for the possibility of individual evaluation or, perhaps more usefully, aggregate ratings to aid teacher and researcher selection and use of web resources informing intervention implementation. It also puts practitioners and researchers in conversation regarding their perceptions of use, and understanding of factors such as credibility, accessibility, and feasibility.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing.* Washington, D.C.: AERA.

Alson, J. (2019). Stress among public school teachers. *Journal of Research Initiatives*, *4*(2), 3.

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., De Freitas, S., Louchart, S., ... & De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology, 46*(2), 391-411.

Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education*, *27*(1), 10-20.

Bangor, A., Kortum, P. & Miller, J.A. (2008). The System Usability Scale (SUS): An empirical evaluation. *International Journal of Human-Computer Interaction, 24(*6), 574-594.

Beahm, L.A., Mandracchia, N.R., Cook, B.A., & Johnson, A.H. (Under Review). How do preservice teachers engage with research-bases websites? Manuscript Under Review.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Q. Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87.

Ben-Zeev, D., Brenner, C. J., Begale, M., Duffecy, J., Mohr, D. C., & Mueser, K. T. (2014). Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia Bulletin, 40*(6), 1244-1253.

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.

Briesch, A.M., Chafouleas, S.M., Neugebager, 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.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, *33*(8), 3-15.

Borko, H., Mayfield, V., Mario, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education*, *13*(3), 259-278.

Brooke, J. (1996). SUS: a "quick and dirty' usability. *Usability Evaluation in Industry*, *189*(3).

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

Buren, M.K., Johnson, A.H., Maggin, D.M., Bains, B.K., Ledoux Galligan, M.R. and Couch, L.K. (2021), "Research Utilization in Special Education", Cook, B.G., Tankersley, M. and Landrum, T.J. (Ed.) *The Next Big Thing in Learning and Behavioral Disabilities* (*Advances in Learning and Behavioral Disabilities, Vol. 31*), Emerald Publishing Limited, Bingley, pp. 29-46.

Carnine, D. (1997). Bridging the research-to-practice gap. *Exceptional Children, 63(4),* 513–521.

Chaffee, R. K., Briesch, A. M., Volpe, R. J., Johnson, A. H., & Dudley, L. (2020). Effects of a class-wide positive peer reporting intervention on middle school student behavior. *Behavioral Disorders*, *45*(4), 224-237.

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., Briesch, A.M., Neugebauer, S. R., & Riley-Tillman, T. C. (2011). Usage Rating Profile – Intervention (Revised). Storrs, CT: University of Connecticut.

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.

Choi, M., Cristol, D., & Gimbert, B. (2018). Teachers as digital citizens: The influence of individual backgrounds, internet use and psychological characteristics on teachers' levels of digital citizenship. *Computers & Education*, *121*, 143-161.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 212–225.

ClassDojo. (n.d.). *About Us*. ClassDojo. Retrieved February 25, 2022, from https://www.classdojo.com/about/

Coburn, C. E., Honig, M. I. and Stein, M. K. (2009a). "What's the evidence on district's use of evidence?". In Educational improvement: What makes it happen and why?, Edited by: Bransford, J., Stipek, D. J., Vye, N. J., Gomez, L. and Lam, D. 67–86. Cambridge, MA: Harvard Educational Press.

Coburn, C. E., Toure, J., & Yamashita, M. (2009b). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *The Teachers College Record, 111*, 1115–1161.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective, 9*(4), 173-206.

Cook, C. R., Frye, M., Slemrod, T., Lyon, A. R., Renshaw, T. L., & Zhang, Y. (2015). An integrated approach to universal prevention: Independent and combined effects of PBIS and SEL on youths' mental health. *School Psychology Quarterly*, *30*(2), 166.

Cramer, K. M. & Castro-Olivo, S. (2015). Effects of a culturally adapted social-emotional learning intervention program on students' mental health. *Contemporary School Psychology, 20(2), 118-129.*

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Sage publications.

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. Journal of Educational Change, 17(1), 7-28.

Education Planner. (n.d.). What's Your learning style? 20 questions. Retrieved February 25, 2022, from http://www.educationplanner.org/students/self-assessments/learning-styles-quiz.shtml

Eggleston, J. (2018). *Teacher decision-making in the classroom.* Routledge.

Feldmann, A., Gasser, O., Lichtblau, F., Pujol, E., Poese, I., Dietzel, C., ... & Smaragdakis, G. (2021, March). Implications of the COVID-19 Pandemic on the

Internet Traffic. In *Broadband Coverage in Germany; 15th ITG-Symposium* (pp. 1-5). VDE.

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

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531-540.

Fogel, J., & Zachariah, S. (2017). Intentions to use the yelp review website and purchase behavior after reading reviews. *Journal of Theoretical and Applied Electronic Commerce Research*, *12*(1), 53-67.

Gilson, C. M., Beach, K. D., & Cleaver, S. L. (2018). Reading motivation of adolescent struggling readers receiving general education support. *Reading & Writing Quarterly*, *34*(6), 505-522.

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

Gueldner, B.A., & Merrell, K.W. (2011). The effectiveness of a social and emotional learning program with middle school students in the general education setting and the effect of consultation on student outcomes. *Journal of Educational and Psychological Consultation, 21*, 1–27.

Gunderson, J., Rangin, H. B., & Hoyt, N. (2006, October). Functional web accessibility techniques and tools from the university of Illinois. In Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (pp. 269-270).

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014, November). Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics* (pp. 228-243). Springer, Cham.

Haertel, E. (2013). Expanding views of interpretation/use arguments. *Measurement: Interdisciplinary Research and Perspectives*, *11*(1–2), 68–70.

Hargittai, E., Piper, A. M., & Morris, M. R. (2019). From internet access to internet skills: digital inequality among older adults. *Universal Access in the Information Society*, *18*(4), 881-890.

Hamdani, B. (2020). Teaching reading through reciprocal teaching method. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, *7*(1), 23-34.

Hester, O. R., Bridges, S. A., & Rollins, L. H. (2020). 'Overworked and underappreciated': special education teachers describe stress and attrition. *Teacher Development*, *24*(3), 348-365.

Howard-Jones, P. A. (2014). Neuroscience and education: myths and messages. *Nature Reviews Neuroscience*, *15*(12), 817-824.

Hunsaker, A., & Hargittai, E. (2018). A review of Internet use among older adults. *New Media & Society*, *20*(10), 3937-3954.

Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use a measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice, 40*(1), 92-100.

International Standard for Organization (2018). Usability of consumer products for public use. Retrieved December 12, 2020 from: https://www.iso.org/obp/ui/#iso:std:iso:ts:20282:-2:ed-2:v1:en

Islam, M. R. (2018). Sample size and its role in Central Limit Theorem (CLT). *Computational and Applied Mathematics Journal*, *4*(1), 1-7.

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.

Jiang, Z., Wang, W., Tan, B. C., & Yu, J. (2016). The determinants and impacts of aesthetics in users' first interaction with websites. *Journal of Management Information Systems, 33*(1), 229-259.

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Kane, M. (2013a). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448-457.

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

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211.

Kelly, B., Sloan, D., Brown, S., Seale, J., Petrie, H., Lauke, P., & Ball, S. (2007, May). Accessibility 2.0: people, policies and processes. In Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A) (pp. 138-147).

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. School Psychology Quarterly, 28, 210 –226.

Kilgus, S. P., Sims, W. A., von der Embse, N. P., & Riley-Tillman, T. C. (2015). Confirmation of models for interpretation and use of the Social and Academic Behavior Risk Screener (SABRS). *School Psychology Quarterly*, *30*(3), 335.

Koehler, M. J., & Mishra, P. (2005). What happens when teachers design educational technology? The development of technological pedagogical content knowledge. *Journal of Educational Computing Research*, *32*(2), 131-152.

Kramer, T.J., Caldarella, P., Young, R., Fischer, L., & Warren, J.S. (2014). Implementing Strong Kids school-wide to reduce internalizing behaviors and increase prosocial behaviors. *Education and Treatment of Children, 37*, 659–680.

Krueger, R. A. (2014). Focus groups: A practical guide for applied research. Sage publications.

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, *70*(2), 144.

Labaree, D. (2018). “An Uneasy Relationship: The History of Teacher Education in the University 1.” In Who Decides Who Becomes a Teacher? (pp. 68-88). Routledge.

Lawrence, D., & Tavakol, S. (2006). Balanced website design: Optimising aesthetics, usability and purpose. Springer Science & Business Media.

Lindsay, B., & Poindexter, M. T. (2003). The Internet: Creating equity through continuous education or perpetuating a digital divide? *Comparative Education Review*, *47*(1), 112-122.

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

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp.com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Marchant, M., Brown, M., Caldarella, P., & Young, E. (2010). Effects of Strong Kids curriculum on students at risk for internalizing disorders: A pilot study. *Journal of Evidence-Based Practices in Schools, 11*(2), 123–143.

McCambridge, J., De Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS one*, *7*(6), e39116.

Madden, A., Ford, N., Miller, D., & Levy, P. (2005). Using the Internet in teaching: The views of practitioners (A survey of the views of secondary school teachers in Sheffield, UK). *British Journal of Educational Technology, 36*(2), 255-280.

Malvik, C. (2020, August 17). *4 types of learning styles: How to accommodate a diverse group of students*. Rasmussen University. Retrieved February 25, 2022, from https://www.rasmussen.edu/degrees/education/blog/types-of-learning-styles/

Mandracchia, N.R. & Sims, W.A. (2020). Development of the Usage Rating Profile-Web Resource (URP-WR): Using assessment to inform web resource selection. *Computers in the Schools, 37(4),* 269-291.

Merrell, K. W., Juskelis, M. P., Tran, O. K., & Buchanan, R. (2008). Social and emotional learning in the classroom: Evaluation of strong kids and strong teens on students' social-emotional knowledge and symptoms. *Journal of Applied School Psychology*, *24*(2), 209-224.

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.

Moss, P. A. (2016). Shifting the focus of validity for test use. Assessment in Education: *Principles, Policy & Practice, 23*(2), 236-251.

Newton, P. M., & Salvi, A. (2020). How common is belief in the learning styles neuromyth, and does it matter? A pragmatic systematic review. *Frontiers in Education, 5*, article 604251.

Okkinga, M., van Steensel, R., van Gelderen, A. J., & Sleegers, P. J. (2018). Effects of reciprocal teaching on reading comprehension of low-achieving adolescents. The importance of specific teacher skills. *Journal of Research in Reading, 41*(1), 20-41.

Opfer, V. D., Kaufman, J. H., & Thompson, L. E. (2016). *Implementation of K–12 state standards for mathematics and English language arts and literacy* [Product Page]. https://www.rand.org/pubs/research_reports/RR1529-1.html

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776.

Orne, M. T. (1996). "Demand characteristics." In Introducing psychological research (pp. 395-401). Palgrave, London.

Palincsar, A. S., & Brown, A. L. (1986). Interactive teaching to promote independent learning from text. *The Reading Teacher*, *39*(8), 771-777.

Payan, A. M., Keller-Margulis, M., Burridge, A. B., McQuillin, S. D., & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention*, *45*(1), 51-64.

Pilten, G. (2016). The evaluation of effectiveness of reciprocal teaching strategies on comprehension of expository texts. *Journal of Education and Training Studies*, *4*(10), 232-247.

Polikoff, M., Dean, J. (2019). The Supplemental Curriculum Bazaar: Is What's Online Any Good? *Thomas B. Fordham Institute*.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4-15.

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.*

Riley-Tillman, T. C., Chafouleas, S. M., Eckert, T. L., & Kelleher, C. (2005). Bridging the gap between research and practice: A framework for building research agendas in school psychology. *Psychology in the Schools*, *42*(5), 459-473.

Rino, J., Bahr, D. L., Larsen, R. A., Sudweeks, R. R., Robinson, J., Everson, K., & Monroe, E. E. (2021). Examining the validity argument of a survey measuring elementary teachers' implementation of standards-based mathematics teaching: An argument-based approach. *Investigations in Mathematics Learning*, *13*(2), 91-106.

Sanetti, L. M. H., & Luh, H. J. (2019). Fidelity of implementation in the field of learning disabilities. *Learning Disability Quarterly*, *42*(4), 204-216.

Sawyer, A. G., & Myers, J. (2018). Seeking comfort: How and why preservice teachers use internet resources for lesson planning. *Journal of Early Childhood Teacher Education, 39*(1), 16-31.

Schrock, K. (2020). "Schrock Guide." Retrieved December 17, 2020 from: https://www.schrockguide.net/critical-evaluation.html

Shelton, C. C., Koehler, M. J., Greenhalgh, S. P., & Carpenter, J. P. (2021). Lifting the veil on TeachersPayTeachers.com: An investigation of educational marketplace offerings and downloads. *Learning, Media and Technology*, 1-20.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 268-280.

Silver, L. (2019, February 5). *Smartphone ownership is growing rapidly around the world, but not always equally*. Pew Research Center's Global Attitudes Project. Retrieved February 22, 2022, from https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/

Tarchi, C., & Pinto, G. (2016). Reciprocal teaching: Analyzing interactive dynamics in the co-construction of a text's meaning. *The Journal of Educational Research, 109*(5), 518-530.

Teaching Resources & Lesson Plans. Teachers Pay Teachers. (n.d.). Retrieved February 25, 2022, from https://www.teacherspayteachers.com/

Teig, N., Scherer, R., & Nilsen, T. (2019). I know I can, but do I have the time? The role of teachers' self-efficacy and perceived time constraints in implementing cognitive-activation strategies in science. *Frontiers in Psychology,* 1697.

Test, D. W., Kemp-Inman, A., Diegelmann, K., Hitt, S. B., & Bethune, L. (2015). Are online sources for identifying evidence-based practices trustworthy? An evaluation. *Exceptional Children*, *82*(1), 58–80.

Tuch, A. N., Bargas-Avila, J. A., & Opwis, K. (2010). Symmetry and aesthetics in website design: It's a man's business. Computers in Human Behavior, 26(6), 1831-1837.

UCONN (2020). Usage Rating Profile Library. Retrieved December 17, 2020 from: https://urp.uconn.edu/library/

U.S. Commission on Civil Rights. (2002, May 17). *Briefing on the consequences of government race data collection bans on civil rights.* https://permanent.fdlp.gov/lps26180/www.usccr.gov/pubs/racedata/summ.htm

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

Vannest, K. J., & Hagan-Burke, S. (2010). Teacher time use in special education. *Remedial and Special Education, 31*(2), 126-142.

Wolf, S. (1974). The real gap between bench and bedside. *New England Journal of Medicine*, *290*(14), 802-803.

Yang, Y. F. (2010). Developing a reciprocal teaching/learning system for college remedial reading instruction. *Computers & Education*, *55*(3), 1193-1201.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., ... & Telzrow, C. (2006). School psychology. A blueprint for training and practice III. Bethesda, MD: National Association of School Psychologists.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., & Telzrow, C. (2008). "The blueprint for training and practice as the basis for best practices." In Best Practices in School Psychology V, 1, 37-70.

Table 1

Demographic Characteristics of Sample by Role

Characteristic		Pre-service Teachers		Doctoral Students		Total	
		n	%	n	%	n	%
Gender							
	Cisgender Woman	35	94%	28	84%	63	90%
	Cisgender Man	2	5%	5	15%	7	10%
Race							
	White	12	32%	24	72%	36	51%
	Black or African American	1	2%	2	6%	3	4%
	American Indian or Alaska Native	1	2%	0	0%	1	1%
	Asian	6	16%	2	6%	8	11%
	Native Hawaiian/ Pacific Islander	1	2%	0	0%	1	1%
	Other	14	37%	5	15%	19	27%
	No Response	2	5%	0	0%	2	2%
Ethnicity							
	Hispanic	24	64%	10	30%	34	48%
	Non-Hispanic	13	35%	23	69%	36	51%

Table 2

Age of Sample by Condition

	Pre-service Teachers			Doctoral Students			Full Sample		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Age	24.62	4.70	20-41	28.09	5.59	22-47	26.33	5.41	20-47

Table 3

Number and Percentage of Total Participants, Doctoral Students, and Pre-service Teachers Who Indicated They Would Use the Web Resource to Help Them Implement the Intervention

	Frequency (%) Total Participants (*n*=70)	Frequency (%) Doctoral Students (*n*=33)	Frequency (%) Pre-Service Teachers (*n*=37)
Scenario A (high on all factors)	66 (94%)	33 (100%)	34 (91%)
Scenario B (low on all factors)	15 (21%)	2 (6%)	13 (35%)
Scenario C (medium on all factors)	28 (40%)	11 (33%)	17 (45%)
Scenario D (high on credibility, low on rest)	26 (31%)	10 (30%)	16 (43%)
Scenario E (low on credibility, high on rest)	21 (30%)	6 (18%)	15 (40%)

Table 4

T-Tests: Pre-service Teacher and Doctoral Students Mean Endorsement of Being Willing to Use the Web Resource to Help Them Implement the Intervention Within Scenarios

Scenario		*n*	Mean	SD	t-test	*p*-value
	Pre-service	37	0.91	0.28		
A	Teachers	29	1.00	0.00	1.78	0.008
	Doctoral Students					
	Pre-service	36	0.35	0.48		
B	Teachers	33	0.06	0.24	-3.22	0.002**
	Doctoral Students					
	Pre-service	36	0.46	0.51		
C	Teachers	29	0.33	0.48	-1.07	0.288
	Doctoral Students					
	Pre-service	36	0.43	0.50		
D	Teachers	29	0.30	0.47	-1.12	0.268
	Doctoral Students					
	Pre-service	37	0.41	0.50		
E	Teachers	30	0.18	0.39	-2.10	0.040*
	Doctoral Students					

*$p < 0.05$. **$p < 0.01$.

Table 5

Number and Percentage of Participants Who Ranked Each Factor as 1-5 for Level of Influence on Decision Making in Each Scenario

Scenario	*n*	Factor	1	2	3	4	5
A	66	Appearance	3 (5%)	6 (9%)	11 (17%)	12 (18%)	34 (52%)
		Accessibility	8 (12%)	21 (32%)	22 (33%)	10 (15%)	5 (8%)
		Feasibility	8 (12%)	19 (29%)	15 (23%)	18 (27%)	6 (9%)
		Credibility	38 (58%)	15 (23%)	9 (14%)	4 (6%)	0 (0%)
		System Support	9 (14%)	5 (8%)	9 (14%)	22 (33%)	21 (32%)
B	67	Appearance	5 (7%)	4 (6%)	10 (15%)	15 (22%)	33 (49%)
		Accessibility	18 (27%)	17 (25%)	20 (30%)	11 (16%)	1 (1%)
		Feasibility	2 (3%)	29 (43%)	19 (28%)	14 (21%)	3 (4%)
		Credibility	40 (60%)	14 (21%)	7 (10%)	3 (4%)	3 (4%)
		System Support	2 (3%)	3 (4%)	11 (16%)	24 (36%)	27 (40%)
C	65	Appearance	5 (8%)	7 (11%)	10 (15%)	8 (12%)	35 (54%)
		Accessibility	16 (25%)	15 (23%)	18 (28%)	13 (20%)	3 (5%)
		Feasibility	8 (12%)	22 (34%)	13 (20%)	17 (26%)	5 (8%)
		Credibility	36 (55%)	18 (28%)	7 (11%)	4 (6%)	0 (0%)
		System Support	0 (0%)	3 (5%)	17 (26%)	23 (35%)	22 (34%)
D	65	Appearance	3 (4%)	9 (14%)	12 (18%)	14 (22%)	27 (42%)
		Accessibility	21 (33%)	17 (26%)	11 (17%)	13 (20%)	3 (5%)
		Feasibility	14 (22%)	24 (37%)	10 (15%)	13 (20%)	4 (6%)
		Credibility	27 (42%)	8 (12%)	14 (22%)	9 (14%)	7 (11%)
		System Support	0 (0%)	7 (11%)	18 (28%)	16 (25%)	24 (37%)
E	67	Appearance	3 (4%)	17 (25%)	10 (15%)	13 (19%)	24 (36%)
		Accessibility	6 (9%)	25 (37%)	25 (37%)	10 (15%)	1 (1%)
		Feasibility	6 (9%)	18 (27%)	11 (16%)	27 (40%)	5 (7%)
		Credibility	52 (78%)	3 (4%)	5 (7%)	3 (4%)	4 (6%)
		System Support	0 (0%)	4 (6%)	16 (24%)	14 (21%)	33 (49%)
Total	330	Appearance	19 (7%)	43 (13%)	53 (16%)	62 (19%)	153 (46%)
		Accessibility	69 (21%)	95 (29%)	96 (29%)	57 (17%)	13 (4%)
		Feasibility	38 (12%)	112 (34%)	68 (21%)	89 (27%)	23 (7%)
		Credibility	193 (58%)	58 (18%)	42 (13%)	23 (7%)	14 (4%)
		System Support	11 (3%)	22 (7%)	71 (22%)	99 (30%)	127 (38%)

Note: Not all participants ranked factors in each Scenario, therefore sample sizes will vary.

Table 6

Number and Percentage of Pre-service Teachers Who Ranked Each Factor as 1-5 for Level of Influence on Decision Making in Each Scenario

Scenario	*n*		1	2	3	4	5
A	37	Appearance	2 (5%)	6 (16%)	7 (19%)	5 (14%)	17 (46%)
		Accessibility	3 (8%)	15 (41%)	11 (30%)	4 (11%)	4 (11%)
		Feasibility	2 (5%)	7 (19%)	9 (24%)	15 (41%)	4 (11%)
		Credibility	23 (62%)	4 (11%)	7 (19%)	3 (8%)	0 (0%)
		System Support	7 (19%)	5 (14%)	3 (8%)	10 (27%)	12 (32%)
B	36	Appearance	2 (6%)	3 (8%)	8 (22%)	5 (14%)	18 (50%)
		Accessibility	11 (31%)	9 (25%)	9 (25%)	7 (19%)	0 (0%)
		Feasibility	0 (0%)	13 (36%)	11 (31%)	9 (25%)	3 (8%)
		Credibility	21 (58%)	8 (22%)	2 (6%)	3 (8%)	2 (6%)
		System Support	2 (6%)	3 (8%)	6 (17%)	12 (33%)	13 (36%)
C	36	Appearance	3 (8%)	5 (14%)	6 (17%)	3 (8%)	19 (53%)
		Accessibility	8 (22%)	12 (33%)	9 (25%)	5 (14%)	2 (6%)
		Feasibility	4 (11%)	8 (22%)	9 (25%)	11 (31%)	4 (11%)
		Credibility	21 (58%)	8 (22%)	4 (11%)	3 (8%)	0 (0%)
		System Support	0 (0%)	3 (8%)	8 (22%)	14 (39%)	11 (31%)
D	36	Appearance	3 (8%)	6 (17%)	5 (14%)	8 (22%)	14 (39%)
		Accessibility	13 (36%)	9 (25%)	6 (17%)	6 (17%)	2 (6%)
		Feasibility	5 (14%)	12 (33%)	7 (19%)	8 (22%)	4 (11%)
		Credibility	15 (42%)	4 (11%)	9 (25%)	6 (17%)	2 (6%)
		System Support	0 (0%)	5 (14%)	9 (25%)	8 (22%)	14 (39%)
E	37	Appearance	2 (5%)	11 (30%)	5 (14%)	7 (19%)	12 (32%)
		Accessibility	5 (14%)	15 (41%)	12 (32%)	5 (14%)	0 (0%)
		Feasibility	2 (5%)	8 (22%)	7 (19%)	17 (46%)	3 (8%)
		Credibility	28 (76%)	1 (3%)	4 (11%)	2 (5%)	2 (5%)
		System Support	0 (0%)	2 (5%)	9 (24%)	6 (16%)	20 (54%)
Total	182	Appearance	12 (7%)	31 (17%)	31 (17%)	28 (15%)	80 (44%)
		Accessibility	40 (22%)	60 (33%)	47 (26%)	27 (15%)	8 (4%)
		Feasibility	13 (7%)	48 (26%)	43 (24%)	60 (33%)	18 (10%)
		Credibility	108 (59%)	25 (14%)	26 (14%)	17 (9%)	6 (3%)
		System Support	9 (5%)	18 (10%)	35 (19%)	50 (27%)	70 (38%)

Note: Not all participants ranked factors in each Scenario, therefore sample sizes will vary.

Table 7

Number and Percentage of Doctoral Students Who Ranked Each Factor as 1-5 for Level of Influence on Decision Making in Each Scenario

Scenario	*n*		1	2	3	4	5
A	29	Appearance	1 (3%)	0 (0%)	4 (14%)	7 (24%)	17 (59%)
		Accessibility	5 (17%)	6 (21%)	11 (38%)	6 (21%)	1 (3%)
		Feasibility	6 (21%)	12 (41%)	6 (21%)	3 (10%)	2 (7%)
		Credibility	15 (52%)	11 (38%)	2 (7%)	1 (3%)	0 (0%)
		System Support	2 (7%)	0 (0%)	6 (21%)	12 (41%)	9 (31%)
B	33	Appearance	3 (10%)	1 (3%)	2 (6%)	10 (32%)	15 (48%)
		Accessibility	7 (23%)	8 (26%)	11 (35%)	4 (13%)	1 (3%)
		Feasibility	2 (6%)	16 (52%)	8 (26%)	5 (16%)	0 (0%)
		Credibility	19 (61%)	6 (19%)	5 (16%)	0 (0%)	1 (3%)
		System Support	0 (0%)	0 (0%)	5 (16%)	12 (39%)	14 (14%)
C	29	Appearance	2 (7%)	2 (7%)	4 (14%)	5 (17%)	16 (55%)
		Accessibility	8 (28%)	3 (10%)	9 (31%)	8 (28%)	1 (3%)
		Feasibility	4 (14%)	14 (48%)	4 (14%)	6 (21%)	1 (3%)
		Credibility	15 (52%)	10 (34%)	3 (10%)	1 (3%)	0 (0%)
		System Support	0 (0%)	0 (0%)	9 (31%)	9 (31%)	11 (38%)
D	29	Appearance	0 (0%)	3 (10%)	7 (24%)	6 (21%)	13 (45%)
		Accessibility	8 (28%)	8 (28%)	5 (17%)	7 (24%)	1 (3%)
		Feasibility	9 (31%)	12 (41%)	3 (10%)	5 (17%)	0 (0%)
		Credibility	12 (41%)	4 (14%)	5 (17%)	3 (10%)	5 (17%)
		System Support	0 (0%)	2 (7%)	9 (31%)	8 (28%)	10 (34%)
E	30	Appearance	1 (3%)	6 (20%)	5 (17%)	6 (20%)	12 (40%)
		Accessibility	1 (3%)	10 (33%)	13 (43%)	5 (17%)	1 (3%)
		Feasibility	4 (13%)	10 (33%)	4 (13%)	10 (33%)	2 (7%)
		Credibility	24 (80%)	2 (7%)	1 (3%)	1 (3%)	2 (7%)
		System Support	0 (0%)	2 (7%)	7 (23%)	8 (27%)	13 (43%)
Total	148	Appearance	7 (5%)	12 (8%)	22 (15%)	34 (23%)	73 (49%)
		Accessibility	29 (20%)	35 (24%)	49 (33%)	30 (20%)	5 (3%)
		Feasibility	25 (17%)	64 (43%)	25 (17%)	29 (20%)	5 (3%)
		Credibility	85 (57%)	33 (22%)	16 (11%)	6 (4%)	8 (5%)
		System Support	2 (1%)	4 (3%)	36 (24%)	49 (33%)	57 (39%)

Note: Not all participants ranked factors in each Scenario, therefore sample sizes will vary.

Figure 1

Percentage of Participants Who Responded "Yes" That They Would Use the Web Resource to Help Them Implement the Intervention in Different Scenarios

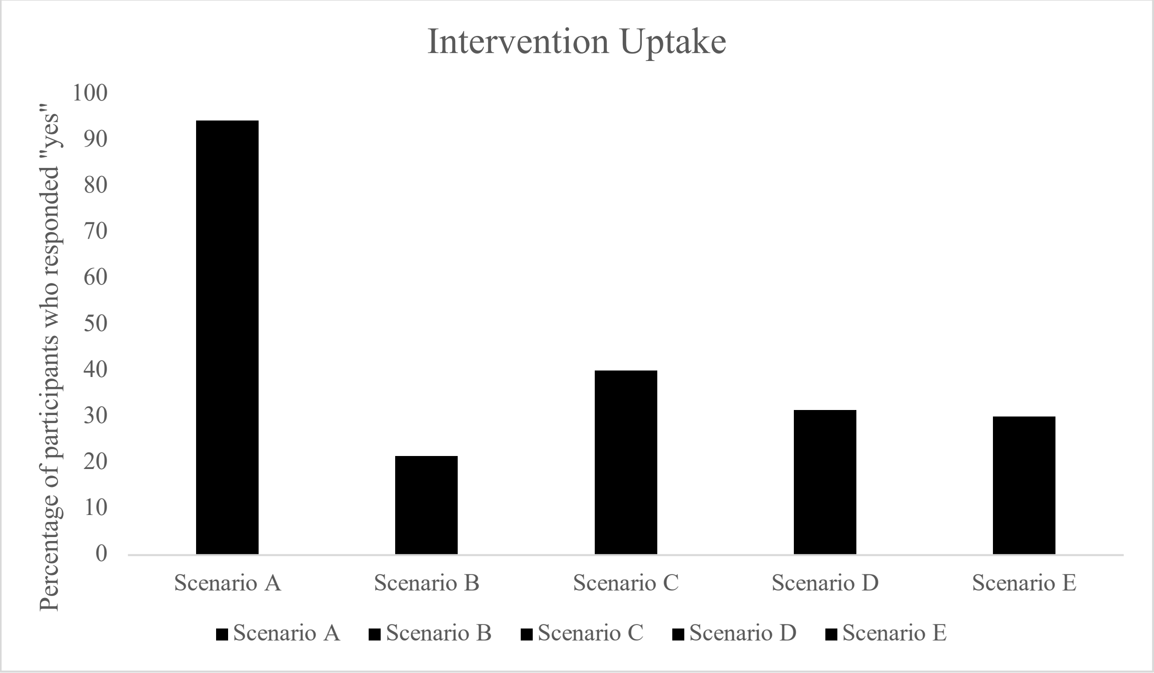


Figure 2

Percentage of Pre-service Teachers and Doctoral Students Who Responded "Yes" That They Would Use the Web Resource to Help Them Implement the Intervention in Different Scenarios

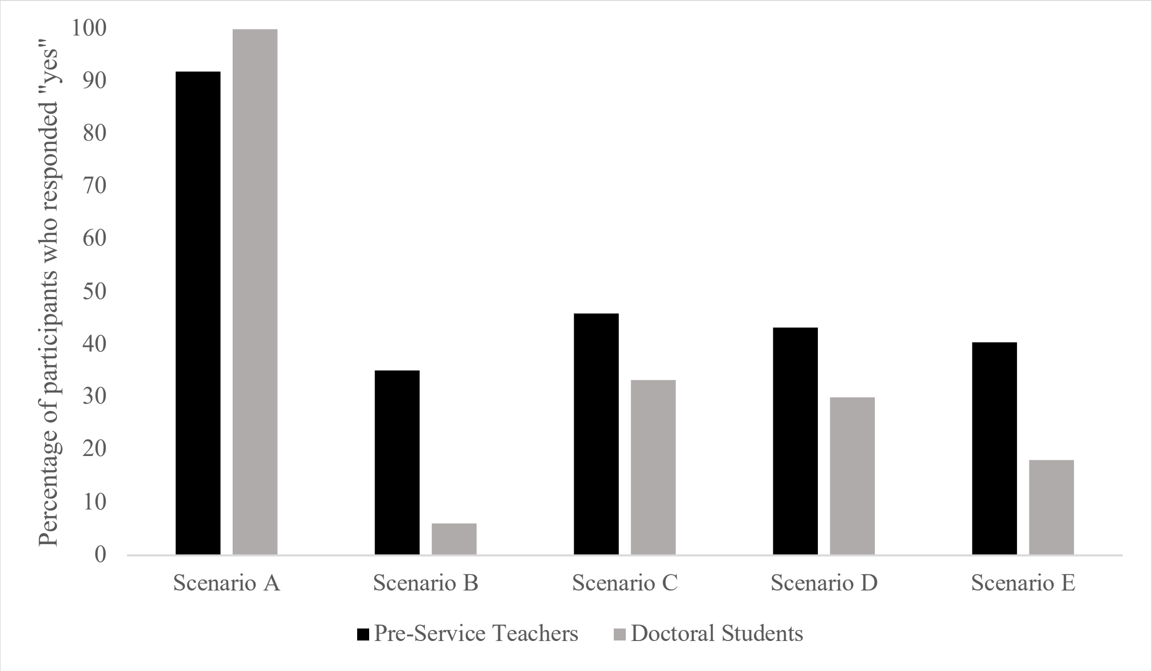


Figure 3

Percentage of Participants Who Ranked Appearance in Each Slot by Scenario

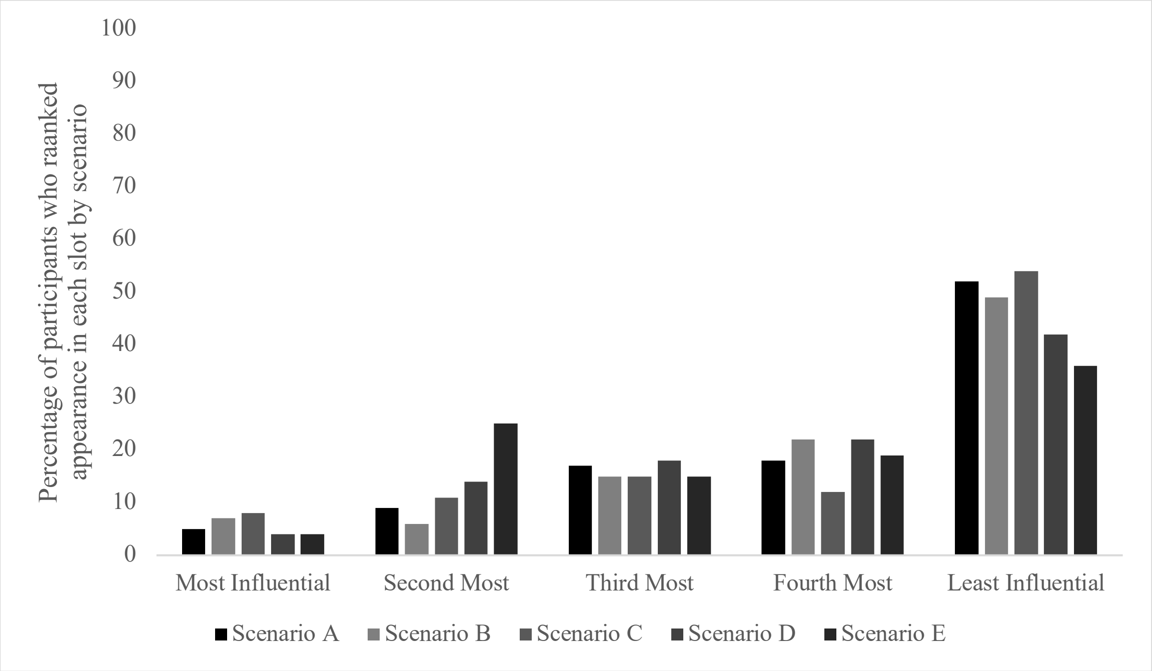


Figure 4

Percentage of Participants Who Ranked Accessibility in Each Slot by Scenario

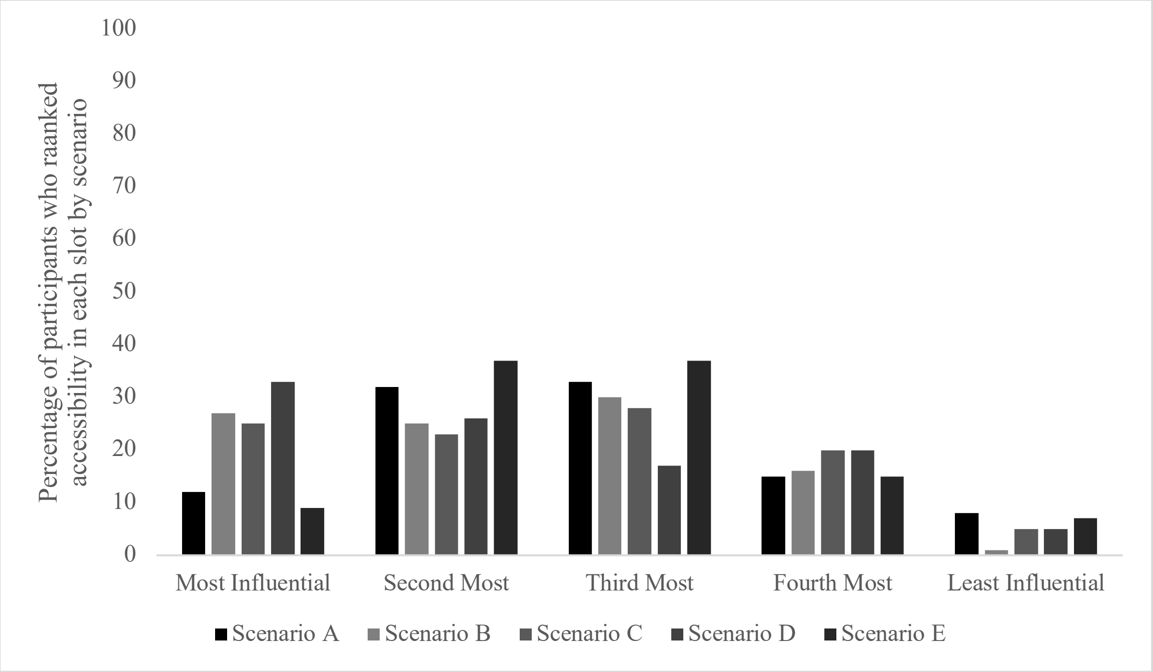


Figure 5

Percentage of Participants Who Ranked Feasibility in Each Slot by Scenario

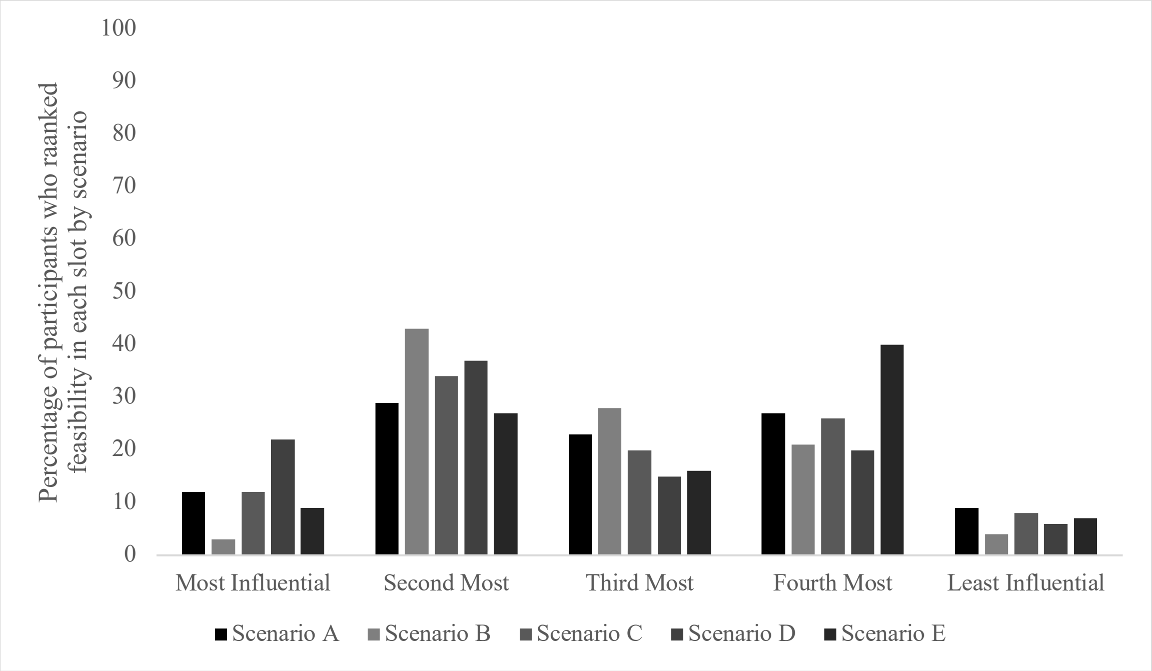


Figure 6

Percentage of Participants Who Ranked Credibility in Each Slot by Scenario

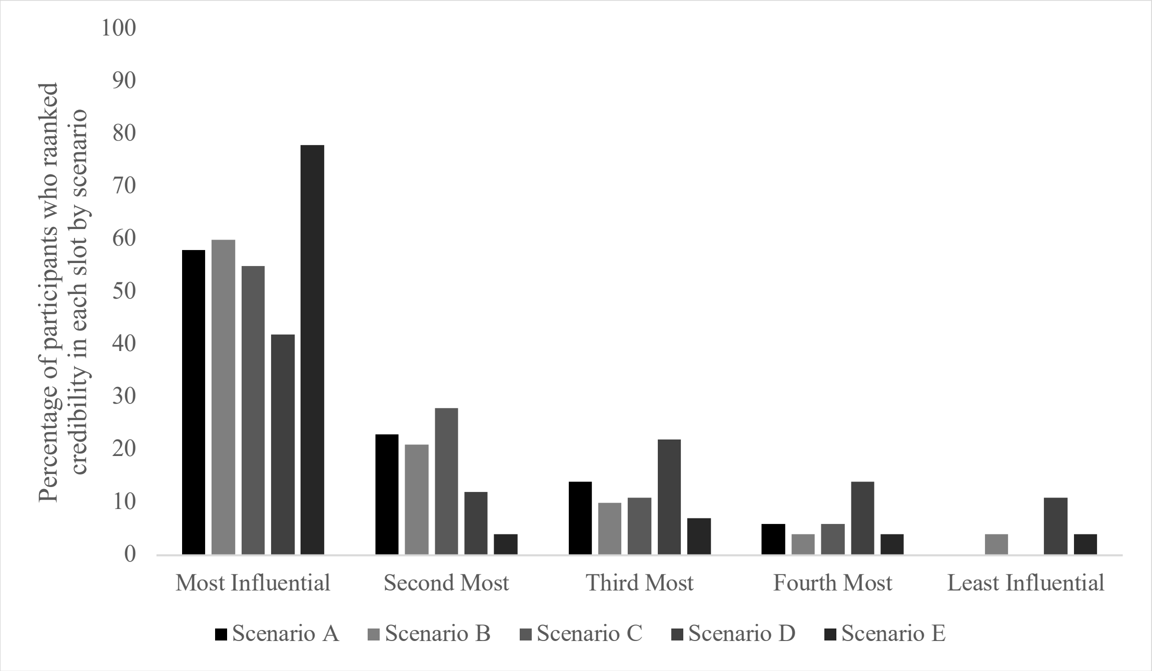


Figure 7

Percentage of Participants Who Ranked System Support in Each Slot by Scenario

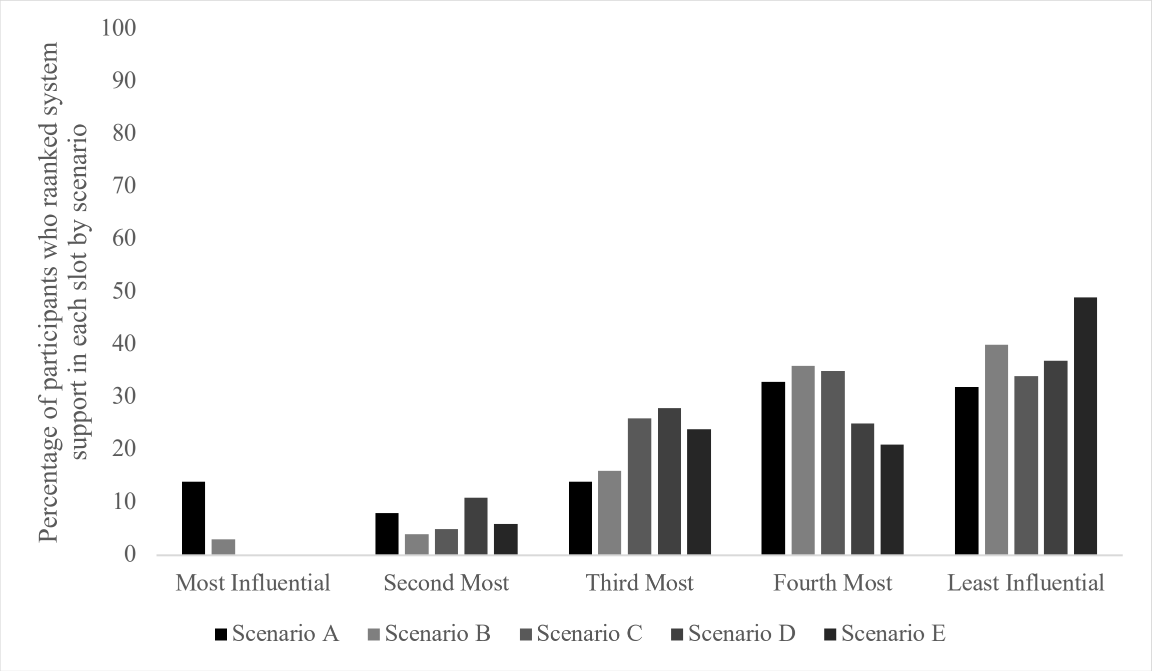


Figure 8

Percentage of Participants Who Ranked Each Factor as Most Influential

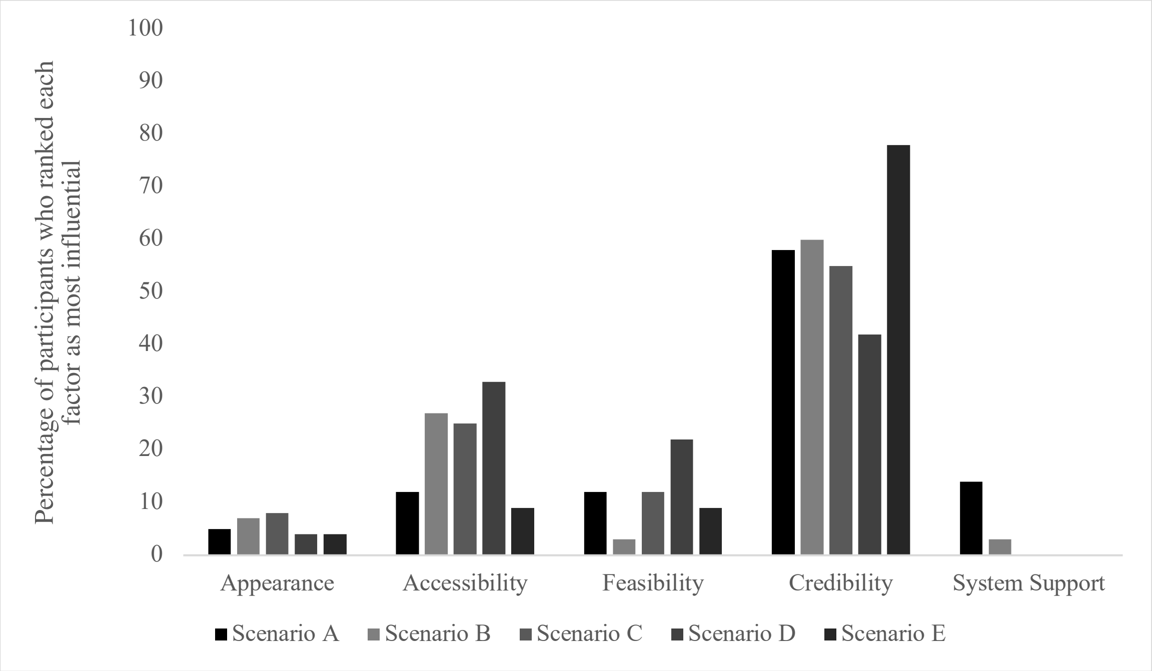


Figure 9

Percentage of Participants Who Ranked Each Factor as Least Influential

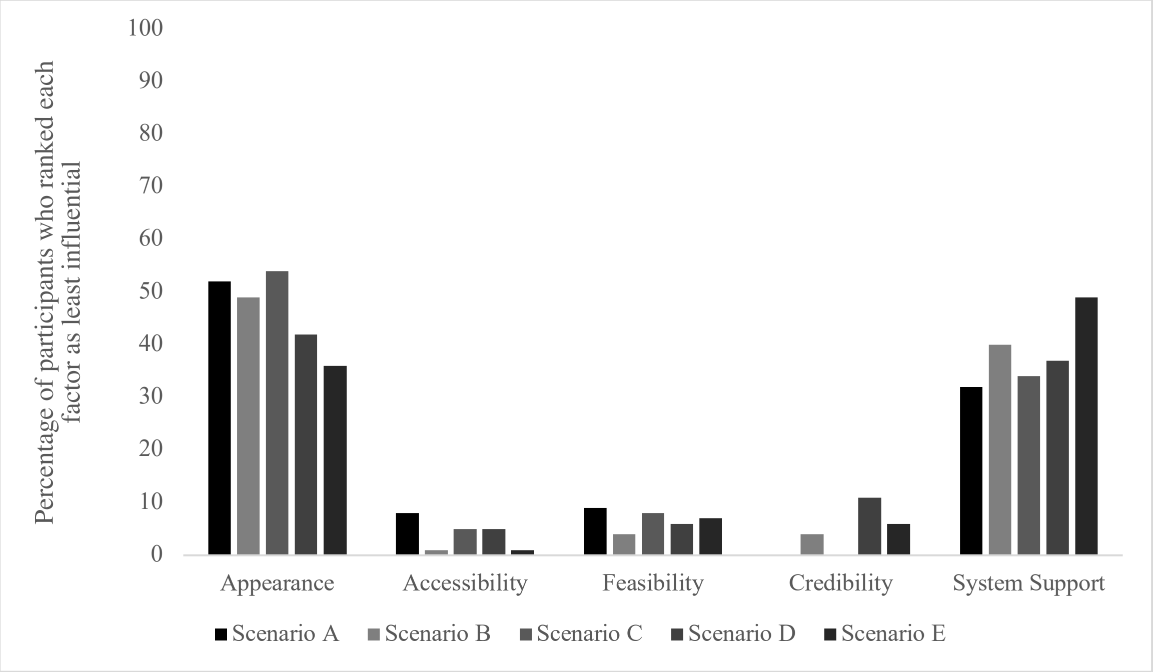


Figure 10

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Appearance in Each Slot by Scenario

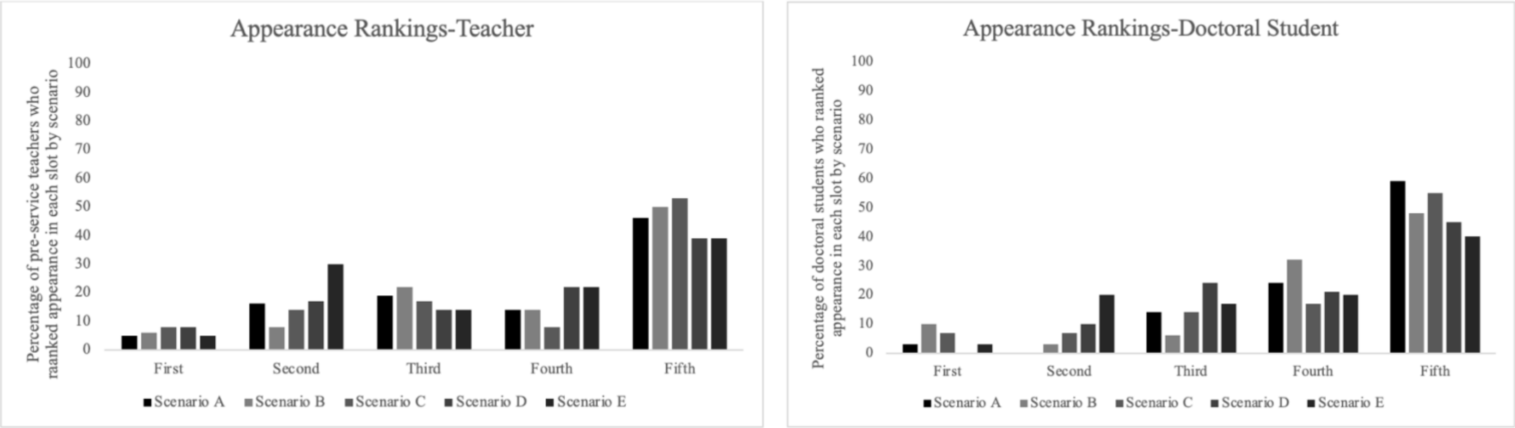


Figure 11

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Accessibility in Each Slot by Scenario

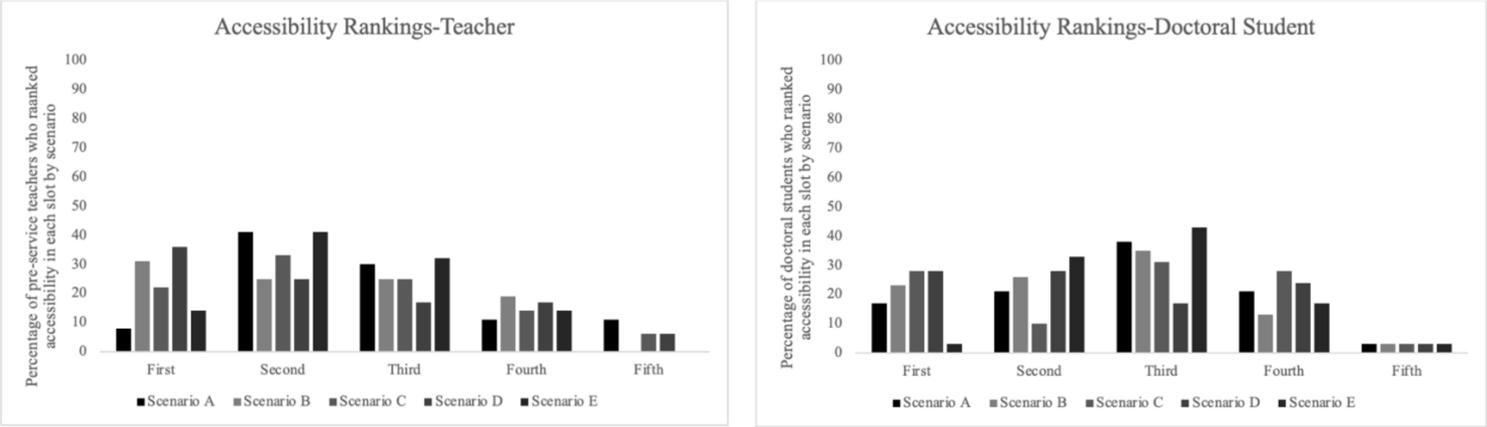


Figure 12

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Credibility in Each Slot by Scenario

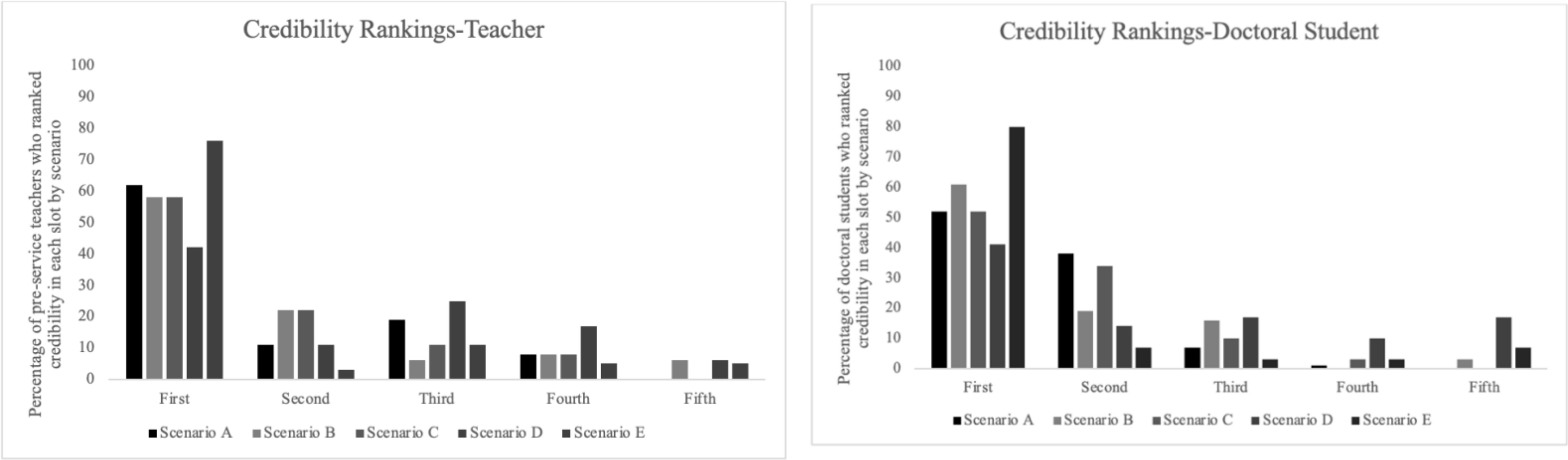


Figure 13

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Feasibility in Each Slot by Scenario

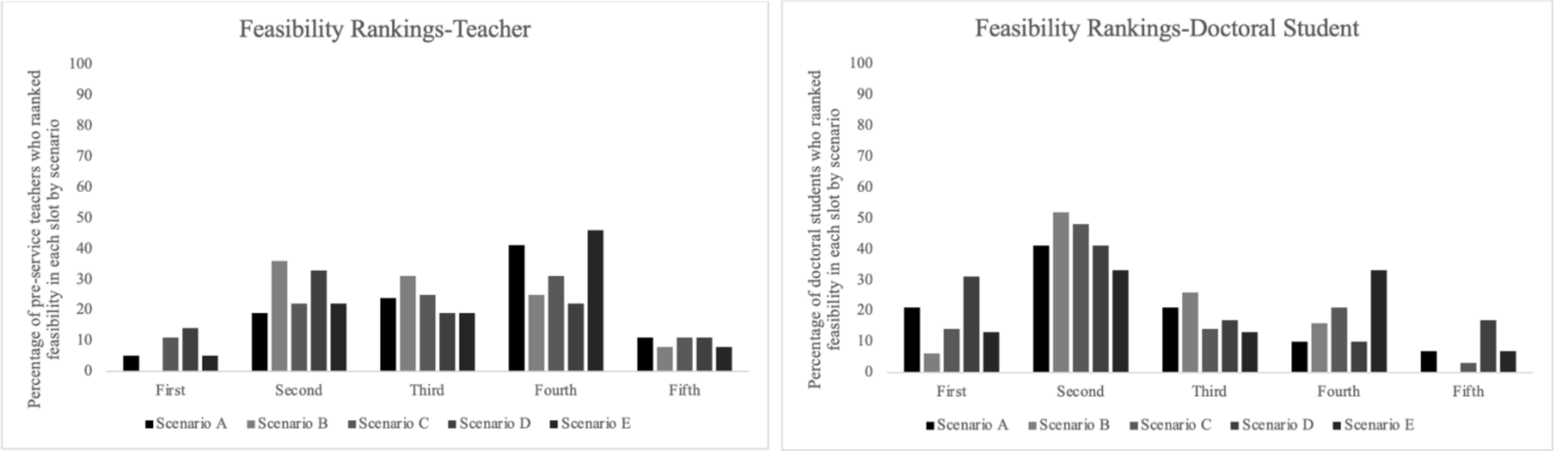


Figure 14

Percentage of Pre-service Teachers and Doctoral Students Who Ranked System Support in Each Slot by Scenario

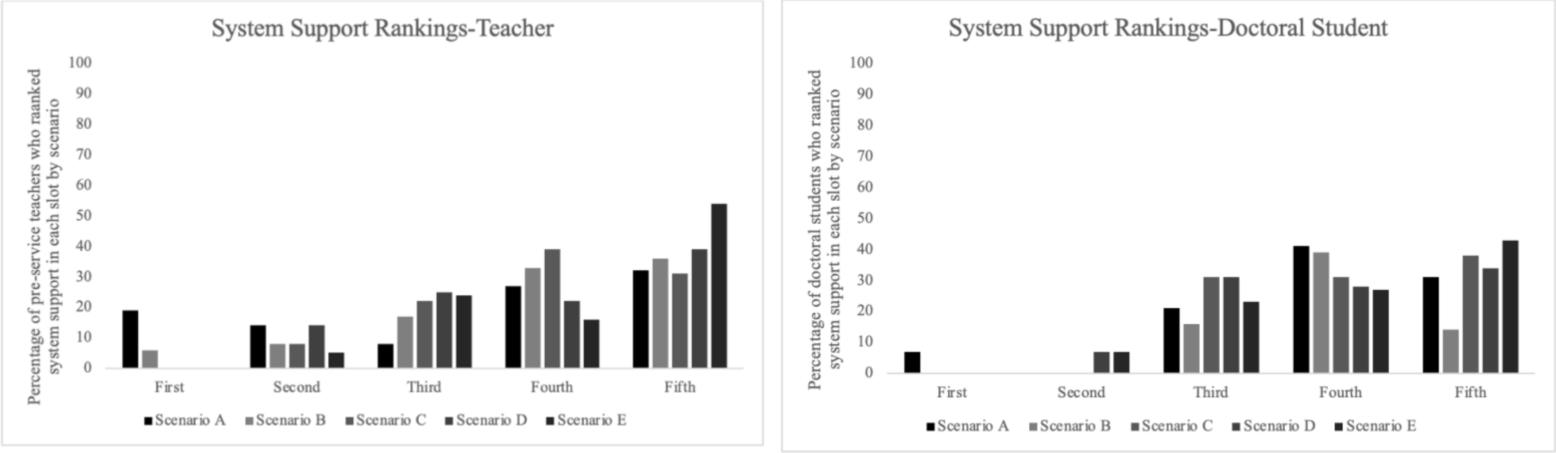


Figure 15

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Each Factor as Most Influential

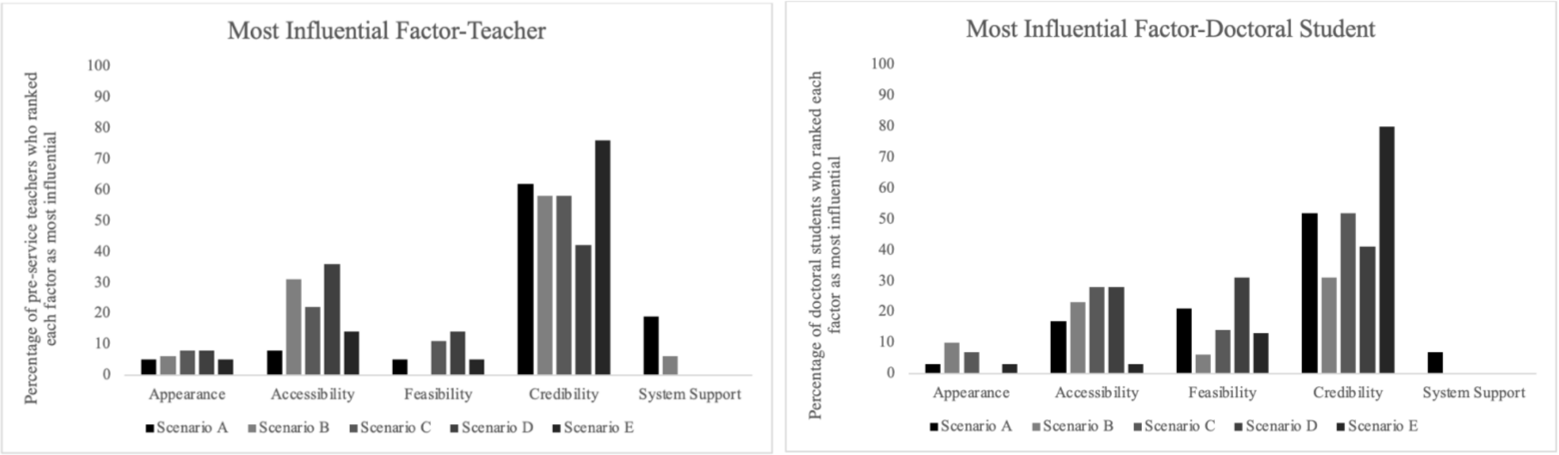
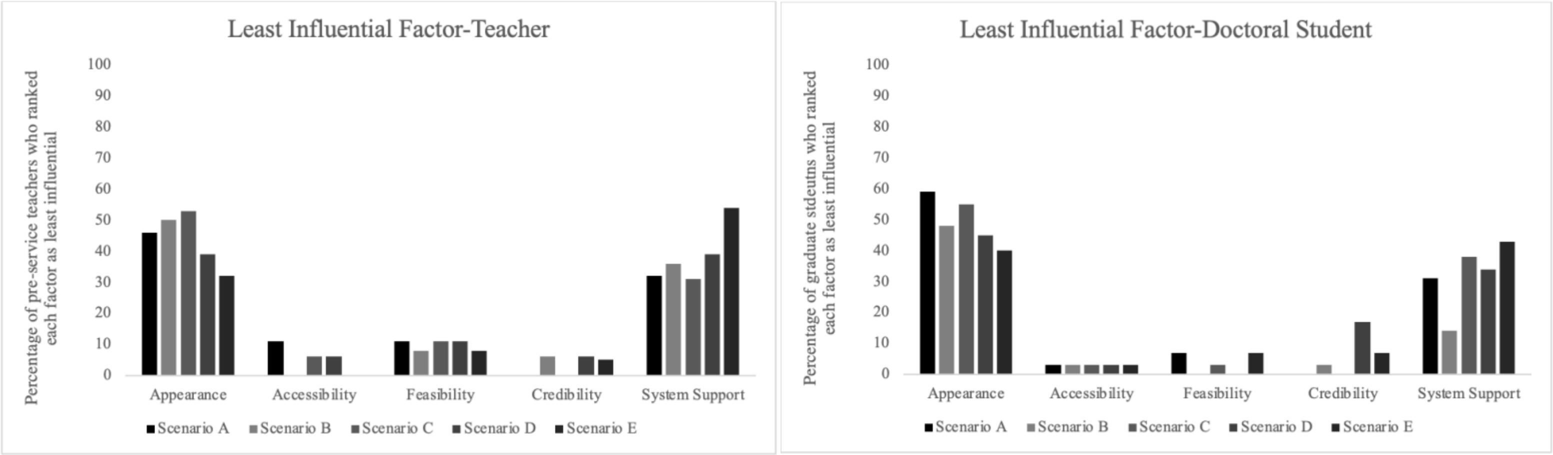


Figure 16

Percentage of Pre-service Teachers and Doctoral Students Who Ranked Each Factor as Least Influential



Appendix A: URP-WR

Items are presented on a 1-6 Likert-type scale from “Strongly Disagree” to “Strongly Agree.”

1. This resource was easy to find.
2. It was difficult to find this resource from a simple Google search.*
3. I could only implement recommendations from this resource with assistance from other adults.**
4. The resource cites its original sources.
5. This resource is aesthetically pleasing.
6. Implementation of the recommendations made in this resource would require support from my co-workers.**
7. It was easy to find this resource from a simple Google search.
8. Topics are successfully addressed, with clearly presented arguments and adequate support to substantiate them.
9. The resource contains all recommendations needed for implementation.
10. Pictures or photographs in the resource add to the information.
11. The resource provides citations from reliable sources.
12. It was easy to find this resource.
13. I would need support from my administrator to implement recommendations made in this resource.**
14. I would know what to say if I were asked how to implement the recommendations provided in this resource.
15. Support from administration would be needed to implement recommendations provided in this resource.**
16. The resource provides citations.
17. The information is from sources known to be reliable.
18. This resource required too many links to find.*
19. This resource looks professional.
20. This resource appropriately represents the context of its cited sources.
21. The sources used by the resource provided appear credible.
22. There is an image map (large clickable graphic with hyperlinks) on the resource.
23. Information for original resource sources is easily identifiable.
24. I understand the components of the recommendations provided in this resource.
25. This resource looks appealing.
26. The resource was updated recently enough for me to trust it.
27. I was able to download this document as a Word doc or PDF for future use.
28. The design of the resource makes me more likely to use it.
29. I wish more resources were designed the way this one is.
30. The site appears well maintained.
31. I believe information from this resource.

* denotes reverse coding (subtract the item score from 6 to obtain the true score)
** denotes potential reverse coding, reverse if the web resource is meant to be selected and implemented independently.

Accessibility: 1, 2*, 7, 12, 18*
Appearance: 5, 10, 19, 22, 25, 26, 27, 28, 29, 30
Plausibility: 4, 8, 9, 11, 14, 16, 17, 20, 21, 23, 24, 31
    Credibility: 4, 8, 11, 16, 17, 20, 21, 23, 31
    Feasibility: 9, 14, 24
System Support: 3**, 6**, 13**, 15**

Scoring guide: The ratings per factor can be calculated by taking a sum or through taking an average of the items in that factor.

Note: For the purpose of this study, factors will be represented by average item scores.

Appendix B: Perceptions of Use

Imagine that you had access to the usability average ratings of a website providing information on a new intervention designed to improve your teaching (graduate level teaching/TA'ing for doctoral students) from 1000 teachers (doctoral students) nationwide. Take between 2 and 5 minutes to review the data provided below, then answer the following questions.

FIRST SCENARIO FOR THAT PARTICIPANT

1. Would you use this website to help you implement this intervention in your practice? YES/NO
    a. Rank how strongly the following factors impacted your decision, with 1 being the most (first consideration) and 5 being the least (last consideration). Please note that the factors are listed in alphabetical order, their current order should not affect your ranking. (RANK)
        i. Credibility
        ii. Feasibility
        iii. Appearance
        iv. Accessibility
        v. System Support

SECOND SCENARIO FOR THAT PARTICIPANT

1. Would you use this website to help you implement this intervention in your practice? YES/NO
    a. Rank how strongly the following factors impacted your decision, with 1 being the most (first consideration) and 5 being the least (last consideration). Please note that the factors are listed in alphabetical order, their current order should not affect your ranking. (RANK)
        i. Credibility
        ii. Feasibility
        iii. Appearance
        iv. Accessibility
        v. System Support

THIRD SCENARIO FOR THAT PARTICIPANT

1. Would you use this website to help you implement this intervention in your practice? YES/NO
    a. Rank how strongly the following factors impacted your decision, with 1 being the most (first consideration) and 5 being the least (last consideration). Please note that the factors are listed in alphabetical order, their current order should not affect your ranking. (RANK)
        i. Credibility
        ii. Feasibility
        iii. Appearance

        iv. Accessibility
        v. System Support

FOURTH SCENARIO FOR THAT PARTICIPANT

1. Would you use this website to help you implement this intervention in your practice? YES/NO
    a. Rank how strongly the following factors impacted your decision, with 1 being the most (first consideration) and 5 being the least (last consideration). Please note that the factors are listed in alphabetical order, their current order should not affect your ranking. (RANK)
        i. Credibility
        ii. Feasibility
        iii. Appearance
        iv. Accessibility
        v. System Support

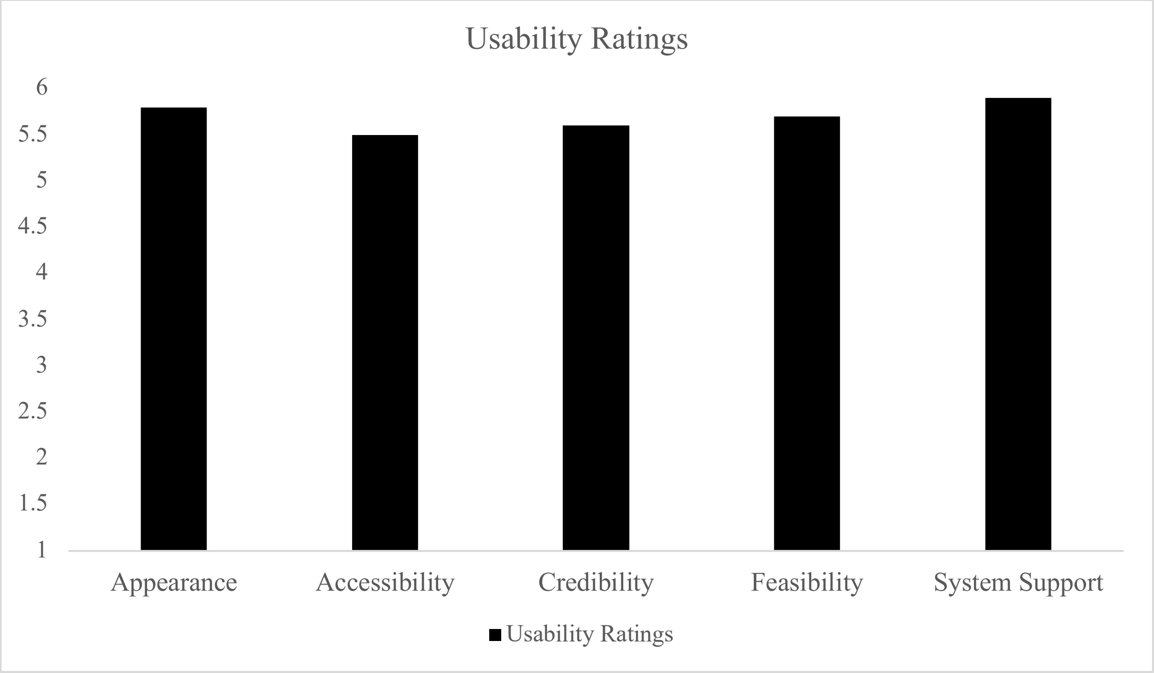
FIFTH SCENARIO FOR THAT PARTICIPANT

1. Would you use this website to help you implement this intervention in your practice? YES/NO
    a. Rank how strongly the following factors impacted your decision, with 1 being the most (first consideration) and 5 being the least (last consideration). Please note that the factors are listed in alphabetical order, their current order should not affect your ranking. (RANK)
        i. Credibility
        ii. Feasibility
        iii. Appearance
        iv. Accessibility
        v. System Support

1. Thinking back to all five scenarios, why did you rank the factors in the way that you did? Please give a brief (1-2 sentences) explanation of your general reasoning. (NARRATIVELY COMPLETED)

Appendix C: Scenarios

Scenario A. Scenario A presented a resource that has been hypothetically rated by 1000 teachers (or doctoral students to match the participant's role) to be high on all five factors of the URP-WR. The presentation appeared as follows.



Appearance 5.8/6.0: the degree to which users perceived this resource to be aesthetically pleasing and thus easy to consume. On ten items related to appearance, 1000 users produced an average item score of 5.8/6.0.

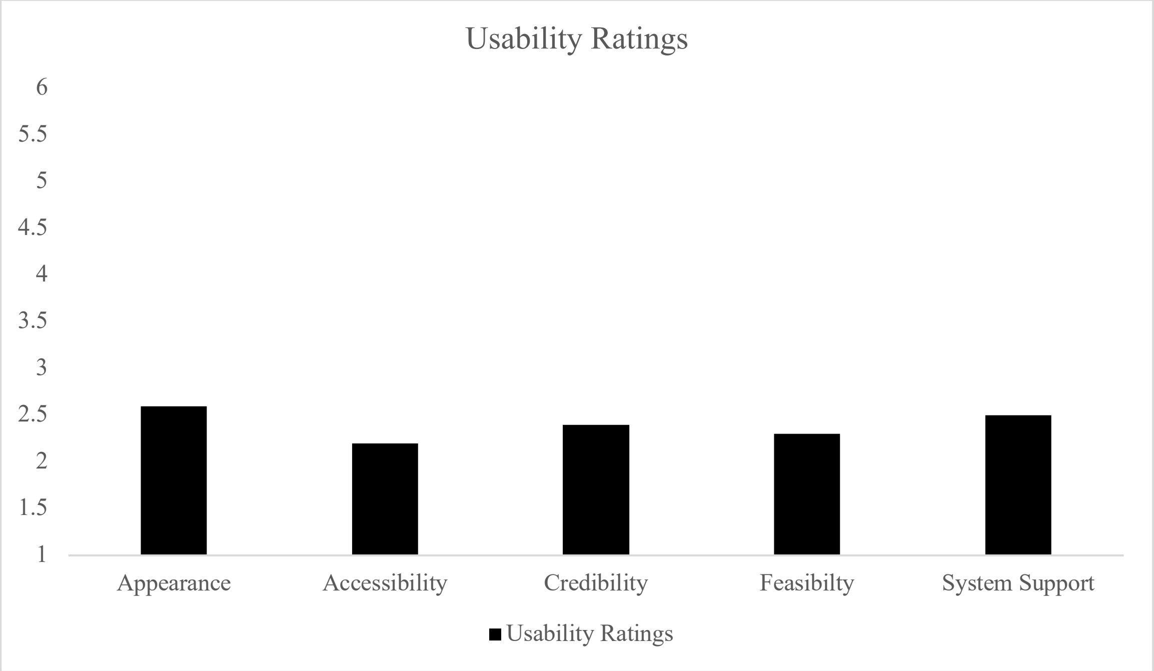
Accessibility 5.5/6.0: the degree to which users perceived this resource to be easy to find and without roadblocks to user-friendliness. On five items related to accessibility, 1000 users produced an average item score of 5.5/6.0.

Credibility 5.6/6.0: the degree to which the users perceived this resource as containing information from credible sources with a solid evidence base. On nine items related to credibility, 1000 users produced an average item score of 5.6/6.0.

Feasibility 5.7/6.0: the degree to which the users perceived this resource as containing information that can be feasibly implemented in a (university-level) classroom setting. On three items related to feasibility, 1000 users produced an average item score of 5.7/6.0.

System Support 5.9/6.0: the degree to which users indicated they would need minimal support from their system (administrators, other teachers, etc.) in order to implement the recommendations. On four items related to system support, 1000 users produced an average item score of 5.9/6.0.

Scenario B. Scenario B presented a resource that has been hypothetically rated by 1000 teachers (or doctoral students) to be low on all five factors of the URP-WR. The presentation appeared as follows.



Appearance 2.6/6.0: the degree to which users perceived this resource to be aesthetically pleasing and thus easy to consume. On ten items related to appearance, 1000 users produced an average item score of 2.6/6.0.

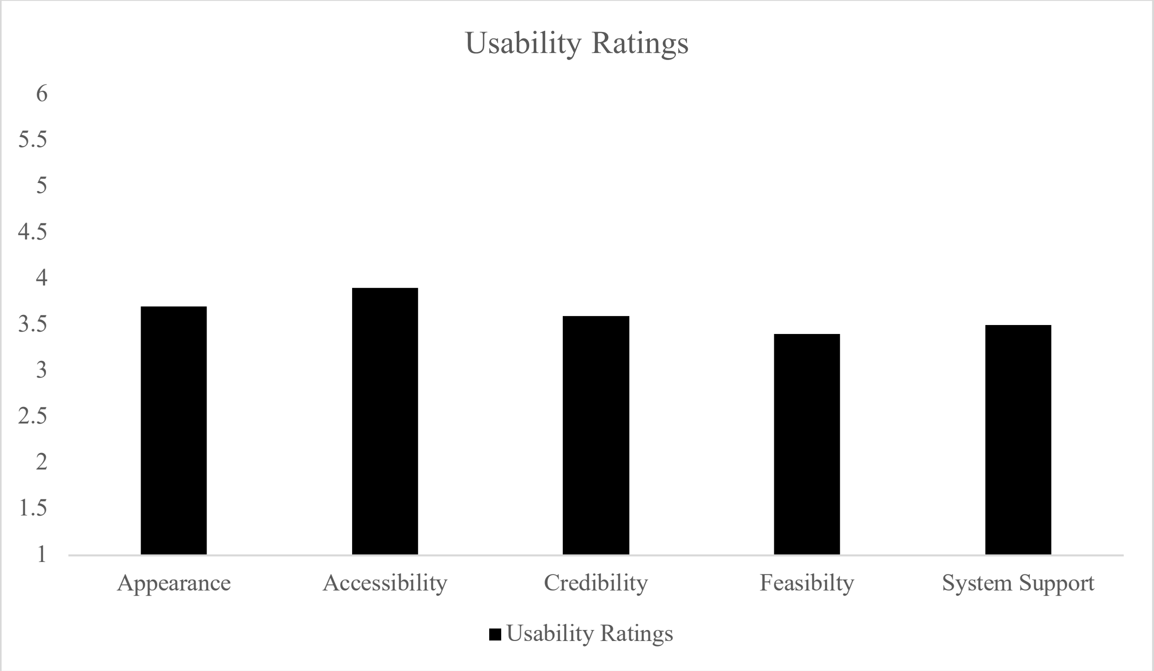
Accessibility 2.2/6.0: the degree to which users perceived this resource to be easy to find and without roadblocks to user-friendliness. On five items related to accessibility, 1000 users produced an average item score of 2.2/6.0.

Credibility 2.4/6.0: the degree to which the users perceived this resource as containing information from credible sources with a solid evidence base. On nine items related to credibility, 1000 users produced an average item score of 2.4/6.0.

Feasibility 2.3/6.0: the degree to which the users perceived this resource as containing information that can be feasibly implemented in a (university-level) classroom setting. On three items related to feasibility, 1000 users produced an average item score of 2.3/6.0.

System Support 2.5/6.0: the degree to which users indicated they would need minimal support from their system (administrators, other teachers, etc.) in order to implement the recommendations. On four items related to system support, 1000 users produced an average item score of 2.5/6.0.

Scenario C. Scenario C presented a resource that has been hypothetically rated by 1000 teachers (or doctoral students) to be medium on all five factors of the URP-WR. The presentation appeared as follows.



<u>Appearance 3.7/6.0</u>: the degree to which users perceived this resource to be aesthetically pleasing and thus easy to consume. On ten items related to appearance, 1000 users produced an average item score of 3.7/6.0.

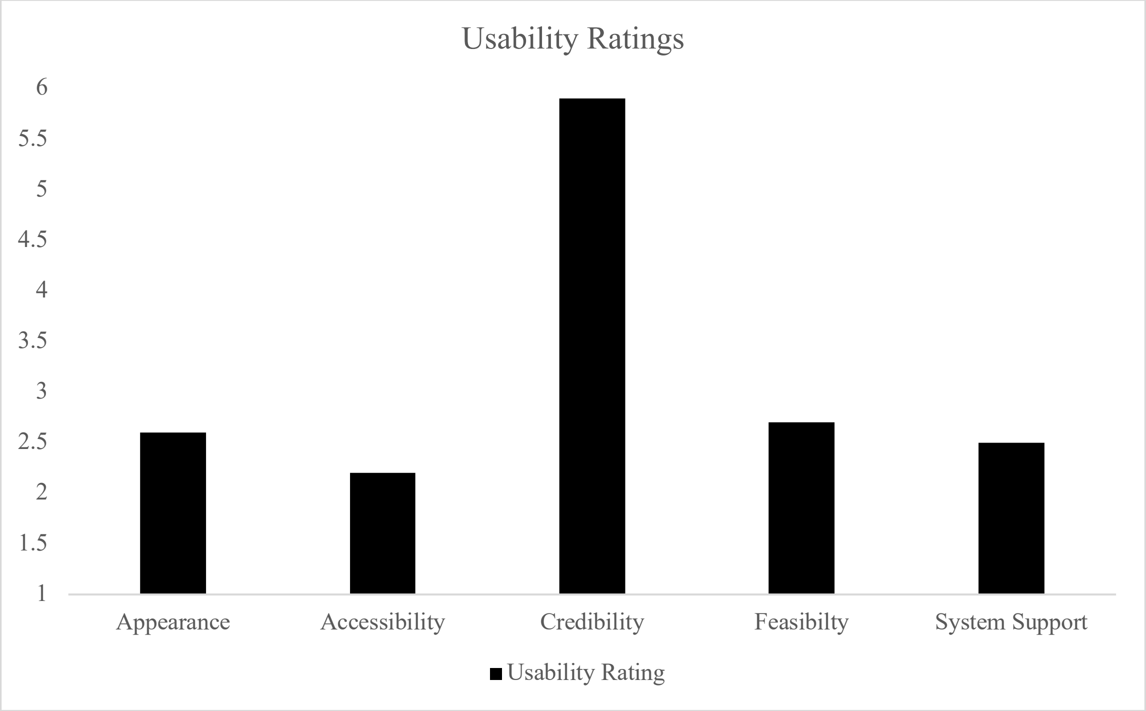
<u>Accessibility 3.9/6.0</u>: the degree to which users perceived this resource to be easy to find and without roadblocks to user-friendliness. On five items related to accessibility, 1000 users produced an average item score of 3.9/6.0.

<u>Credibility 3.6/6.0</u>: the degree to which the users perceived this resource as containing information from credible sources with a solid evidence base. On nine items related to credibility, 1000 users produced an average item score of 3.6/6.0.

Feasibility 3.4/6.0: the degree to which the users perceived this resource as containing information that can be feasibly implemented in a (university-level) classroom setting. On three items related to feasibility, 1000 users produced an average item score of 3.4/6.0.

System Support 3.5/6.0: the degree to which users indicated they would need minimal support from their system (administrators, other teachers, etc.) in order to implement the recommendations. On four items related to system support, 1000 users produced an average item score of 3.5/6.0.

Scenario D. Scenario D presented a resource that has been hypothetically rated by 1000 teachers (or doctoral students) to be high on the credibility factor, but low on the other factors of the URP-WR. The presentation appeared as follows.



Appearance 2.6/6.0: the degree to which users perceived this resource to be aesthetically pleasing and thus easy to consume. On ten items related to appearance, 1000 users produced an average item score of 2.6/6.0.

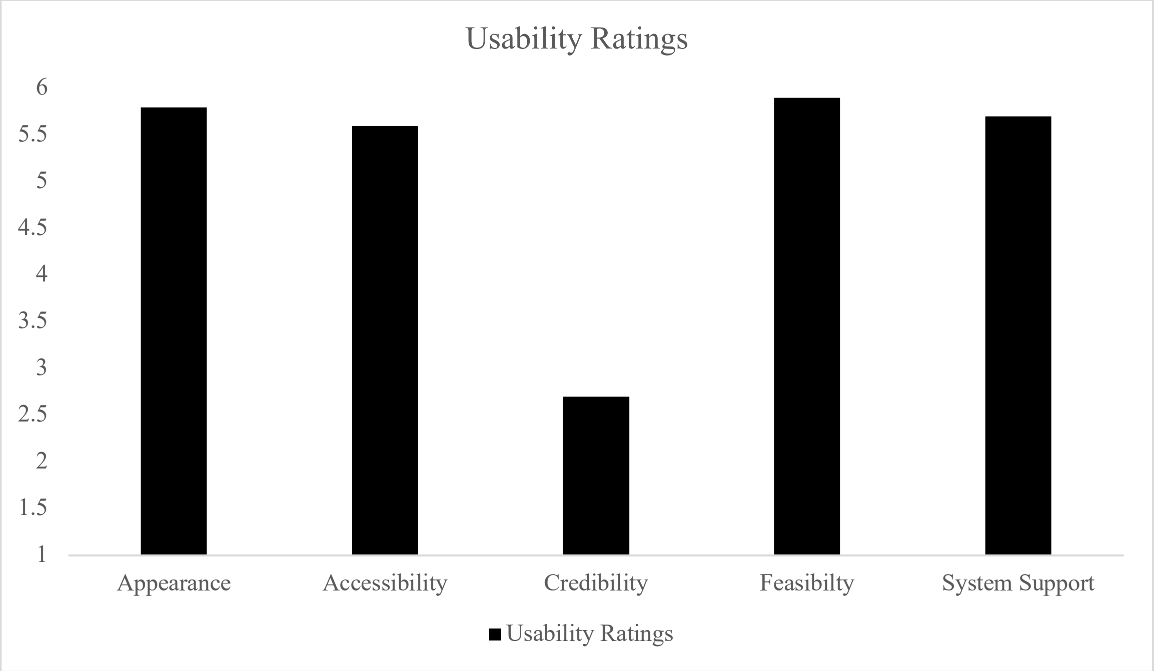
Accessibility 2.2/6.0: the degree to which users perceived this resource to be easy to find and without roadblocks to user-friendliness. On five items related to accessibility, 1000 users produced an average item score of 2.2/6.0.

Credibility 5.9/6.0: the degree to which the users perceived this resource as containing information from credible sources with a solid evidence base. On nine items related to credibility, 1000 users produced an average item score of 5.9/6.0.

Feasibility 2.7/6.0: the degree to which the users perceived this resource as containing information that can be feasibly implemented in a (university-level) classroom setting. On three items related to feasibility, 1000 users produced an average item score of 2.7/6.0.

System Support 2.5/6.0: the degree to which users indicated they would need minimal support from their system (administrators, other teachers, etc.) in order to implement the recommendations. On four items related to system support, 1000 users produced an average item score of 2.5/6.0.

Scenario E. Scenario E presented a resource that has been hypothetically rated by 1000 teachers (or doctoral students) to be low on the credibility factor and high on the other factors of the URP-WR. The presentation appeared as follows.



<u>Appearance 5.8/6.0</u>: the degree to which users perceived this resource to be aesthetically pleasing and thus easy to consume. On ten items related to appearance, 1000 users produced an average item score of 5.8/6.0.

<u>Accessibility 5.5/6.0</u>: the degree to which users perceived this resource to be easy to find and without roadblocks to user-friendliness. On five items related to accessibility, 1000 users produced an average item score of 5.5/6.0.

<u>Credibility 2.7/6.0</u>: the degree to which the users perceived this resource as containing information from credible sources with a solid evidence base. On nine items related to credibility, 1000 users produced an average item score of 2.7/6.0.

Feasibility 5.9/6.0: the degree to which the users perceived this resource as containing information that can be feasibly implemented in a (university-level) classroom setting. On three items related to feasibility, 1000 users produced an average item score of 5.9/6.0.

System Support 5.7/6.0: the degree to which users indicated they would need minimal support from their system (administrators, other teachers, etc.) in order to implement the recommendations. On four items related to system support, 1000 users produced an average item score of 5.7/6.0.