Learning Disabilities: A Contemporary Journal 15(2), 193-207, 2017

Initial Validation of the Usage Rating Profile-Assessment for Use Within German Language Schools

Amy M. Briesch
Northeastern University; USA

Gino Casale
University of Cologne, Germany

Michael Grosche
University of Wuppertal, Germany

Robert J. Volpe
Northeastern University, USA

Thomas Hennemann
University of Cologne, Germany

Modern attempts to explain why some assessment tools are readily adopted by school-based personnel whereas others are not have focused on the concept of usability. Usability encompasses not only the degree to which consumers find an assessment tool to be acceptable, but also the degree to which it is well-understood, believed to be feasible, consistent with local norms, and supported within the larger school environment. The purpose of the current study was to conduct an initial validation of a German-language version of the Usage Rating Profile-Assessment (URP-A), a measure designed to assess the multiple influences on assessment usage. Participants included 101 1st-through 6th-grade teachers in Western Germany. Although findings from an exploratory factor analysis of URP-A items differed somewhat from results found for the original English-language measure, results of the current study suggest that the German URP-A may actually be used to reliably assess multiple dimensions of usability with a fewer number of items.

Keywords: **Assessment, Treatment Usage, Acceptability, Factor Analysis, Teacher SELF-Report**

Introduction

Significance of Data-Informed Decision Making Processes in Preventing Learning Disabilities

The early and systematic identification of learning problems in students is a key element of proactive approaches for the prevention of learning disabilities (Fuchs, 2003). For instance, multi-tiered systems of support (MTSS), such as Response-to-Intervention (RtI), provide a conceptual framework for the early identification of students who struggle with the academic requirements in schools (Grosche & Volpe, 2013). Two of the driving assumptions behind a successful MTSS model are that (a)

*Please send correspondence to: Amy M. Briesch, PhD, Department of Applied Psychology, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA, Phone: 1-617-373-8291, Email: a.briesch@northeastern.edu.

students receive instruction or intervention that is evidence-based, and (b) the intensity of prevention and intervention efforts provided to students are informed by the results of evidence-based assessment tools (e.g., screening, progress monitoring measures) (Eagle, Dowd-Eagle, Snyder, & Holtzman, 2015). Unfortunately, however, although our accumulated base of knowledge regarding both evidence-based programs and assessment has grown substantially in recent years with the development of comprehensive databases such as the What Works Clearinghouse (http://ies.ed.gov/ncee/wwc) and National Center on Intensive Intervention (http://www.intensiveintervention.org), the technologies that we know to produce positive student outcomes are not necessarily being utilized in everyday school settings (Briesch, Chafouleas, Neugebauer, & Riley-Tillman, 2012). The problem is that the assumed effectiveness of the technology is all too often counterbalanced by the practical barriers that exist in the translation of research to practice.

Factors Hypothesized to Influence Applied Usage

Over the past four decades, researchers have sought to better understand those factors that may help explain why some intervention and assessment technologies are embraced by users whereas others are not. An important line of inquiry into this issue was initiated in the early 1980s by Kazdin (1980), who focused on the construct of treatment acceptability. Treatment acceptability was defined as the degree to which prospective users of a treatment believe it to be something that is appropriate, fair, and reasonable for the given problem, and the idea was that users are more likely to put into actual practice those treatments that they find to be acceptable (Kazdin, 1980). Within the field of education, much work was conducted throughout the 1980s in order to better understand the degree to which treatments ranging from math interventions (e.g., Logan & Skinner, 1998) to pharmacological treatments (e.g., Power, Hess, & Bennett, 1995) were believed to be acceptable to teachers, students, and parents, as well as which factors had the greatest influence on perceived acceptability. Results of this line of research suggested that the most acceptable treatments were those that were both effective (i.e. resulting in positive change with minimal side effects) and feasible (i.e. requiring minimal time and resources) for individuals to implement (Reimers, Wacker, & Koeppl, 1987).

Although acceptability continues to be considered to be an important determinant of actual usage, over time researchers have acknowledged the need to consider additional factors beyond acceptability alone. One reason for this expanded consideration has been the fact that some research has found low correlations between how acceptable a treatment is perceived to be and the degree to which it is actually implemented (e.g., Sterling-Turner & Watson, 2002; Mautone, DuPaul, Jitendra, Tresco, Junod, & Volpe, 2009). As a result, more modern ecological conceptualizations of treatment usage have come to incorporate factors believed to influence usage at multiple levels. At the primary level are the implementer-level factors that exist within an individual. Personal acceptability is one important factor at this level; however, an individual's understanding of what the treatment is and how it is intended to be used may also have a notable effect on actual implementation (Reimers et al., 1987; Witt, Noell, LaFleur, & Mortensen, 1997). At the next level are the intervention-level factors that relate to features of the intervention or assessment technology itself. For

example, the extent to which a procedure requires extensive amounts of time or resources will influence the degree to which it is implemented (e.g., Perpletchikova & Kazdin, 2005). Related, those procedures that result in significant disruption to regular classroom activities will tend to be viewed less favorably (Reimers et al., 1987; Witt, 1986). Finally, it is important to consider those broader environmental factors that may influence local usage. That is, even if an intervention or assessment technology is perceived positively by an individual implementer, there may be administrative or philosophical hurdles to implementation within the broader school context. These contextual considerations include the degree to which there is administrative and peer support for the practice, both philosophically and practically speaking (Broughton & Hester, 1993; Buston, Wight, Hart, & Scott, 2002).

Development of the Usage Rating Profile

With the evolution of multi-dimensional conceptualizations of treatment usage came the need for a measure that would be capable of simultaneously assessing the multiple factors that are believed to influence intervention usage. The Usage Rating Profile-Intervention (URP-I; Chafouleas, Briesch, Riley-Tillman, & McCoach, 2009) was therefore developed in order to serve this purpose. Most recent work on the URP-I has supported a six-factor model of usage that considers (a) how acceptable (i.e. Acceptability) and feasible (i.e. Feasibility) a user perceives the intervention to be, (b) the degree to which the intervention is well-understood (i.e. Understanding) and family support is needed for implementation (i.e. Home-School Collaboration), (c) and the degree to which both practical (i.e. System Support) and philosophical (i.e. System Climate) system-level supports are needed (Briesch et al., 2012).

Although the URP was originally designed and validated for use when considering school-based interventions, the tool has recently been extended to consider use of assessment tools as well. Miller, Neugebauer, Chafouleas, Briesch, and Riley-Tillman (2012) adapted the item wording from the URP-I to reflect perceptions of assessment rather than intervention tools (e.g., *I would need additional resources to carry out this assessment* as opposed to *I would need additional resources to carry out this intervention*), thereby creating the Usage Rating Profile-Assessment (URP-A). These researchers then asked 283 public school teachers to complete the URP-A with regard to a teacher-completed behavioral assessment measure (i.e. Direct Behavior Rating). Results of this study suggested that the factor structure of the URP-A was consistent with the six-factor structure of the URP-I; however, the reliabilities of the System Support and Climate scales were found to be lower than in the previous investigation.

The emerging literature base has suggested great promise for use of the URP within school settings; however, to date, research related to this tool has focused exclusively on its use in English-language contexts. In order to ensure comparability in cross-cultural research, however, it is important to make the tool available in other countries (Ziegler & Bensch, 2013). If the URP is to be used by researchers and practitioners in non-English speaking countries, it is necessary not only to translate the measure into the local language but also to verify that the psychometric properties of the measure are similarly strong. The purpose of the current study was therefore to conduct a validation of a German-language version of the URP-A in order to determine whether the factor structure would be consistent (i.e. construct validity) and

whether the German URP-A would allow for sufficiently reliable measurement of the hypothesized factors (i.e. internal consistency reliability).

Method

Participants

Participants included 101 1st through 6th grade teachers (94% female) in a suburban region of Western Germany, who responded to a call for participation in the study. In total, teachers were recruited from 13 elementary schools, four secondary schools, and one special education school. In total, the age of the participating teachers ranged from 26 to 63 (M = 43.10, SD = 10.48) and they had an average of 15.39 years of teaching experience (SD = 9.31; Range = 2-39). A detailed summary of demographic information for these teacher participants is provided in Table 1. The ratings in this study were completed for a sample of 1010 students (39.6% female). The age of the students ranged between 5 and 14 years (M = 8.14, SD = 1.77) and the mean grade level was 2.94 (SD = 1.48).

Table 1. ***Teacher Demographics***

		N	Percentage
Gender			
	Male	6	6%
	Female	95	94%
Age			
	26-35	28	28%
	36-45	34	34%
	46-55	20	20%
	56-65	19	19%
Years Experience			
	1-10	34	34%
	11-20	42	42%
	21-30	16	16%
	31-40	8	8%
	Unknown	1	1%
Training			
	Primary	83	82%
	Secondary	5	5%
	Special Education	6	6%
	Other	6	6%

Procedure

Data were collected between September and November in 2015 as part of a larger study designed to examine the psychometric characteristics of a novel multiple-gated screening measure designed to link screening assessment to the design of classroom-based intervention (i.e. Integrated Teacher Report Form, ITRF; Volpe & Fabiano, 2013). As part of this larger study, all participants were asked to nominate five students in their classroom who struggled with problematic classroom behavior. The researchers then selected an additional five students who were not nominated by the classroom teacher to serve as typical comparison peers. All teacher participants were then asked to complete a packet of rating forms for each of the 10 identified students (i.e. 5 nominated, 5 not nominated). As an incentive for participation, all teachers were entered into a drawing to receive one of two 50€gift cards to a teaching material trade company.

All teacher participants received a packet consisting of (a) a demographic questionnaire, (b) an explanation of rating procedures, and (c) the German language version of the ITRF (ITRF-G). In addition, each participant was asked to complete a second behavioral screening tool (see descriptions in the Measures section below), which was randomly assigned. Finally, teachers were asked to complete the translated version of the URP-A (see Measures below) with regard to each of the screening measures completed.

Although all participants completed the URP-A in response to the ITRF-G and one additional screening measure, responses were selected for a single measure for the purposes of this study. In order to ensure some variability in perception of usability across respondents, data were purposively selected to reflect the range of screening assessment options. That is, the data set was divided into thirds, such that the number of responses based on each of the three screening measures was roughly equivalent (ITRF: n = 40, 39.6 %; LSL: n = 27, 26.7 %; SDQ: n = 34, 33.7 %).

Measures

Integrated Teacher Report Form-German language version (ITRF-G). The ITRF uses a multiple-gated approach to proactively identify those students who might benefit from additional behavioral supports in the classroom. Teachers are first asked to complete a brief 16-item version of the ITRF to rate the degree to which their students' behavior interferes with their own learning or the learning of others. Next, teachers complete a 43-item rating scale for the five students receiving the highest brief ITRF scores, which asks respondents to indicate the degree to which particular behaviors are of concern for the student using a 3-point scale (i.e. slight concern, moderate concern, strong concern). A total score is then calculated for each student so that individual students may be prioritized for follow-up intervention. Unlike many other behavioral screening tools, which focus on identifying underlying indicators of psychopathology, the items on the ITRF represent behaviors that have the potential to impair classroom functioning, but which are believed to be malleable targets of classroom intervention.

Table 2. ***Pattern Coefficients and Communalities for the Usage Rating Profile-Assessment***

Item		Factor I	Factor II	Factor III	Factor IV	Factor V	
		P	P	P	P	P	h2
2	I would need additional resources to carry out this assessment.	**-.57**					.37
3	I would be able to allocate my time to implement this assessment.	**.65**					.56
8	The total time required to implement the assessment procedure would be manageable.	**.74**					.73
18	This assessment is too complex to carry out accurately.	**-.65**					.54
19	These assessment procedures are consistent with the way things are done in my system.	**.70**					.71
26	The amount of time required for record keeping would be reasonable.	**.74**					.72
5	A positive home-school relationship is needed to use this assessment.		**.78**				.69
15	Parental collaboration is required in order to use this assessment.		**.81**				.65
27	Regular home-school communication is needed to implement these assessment procedures.		**.88**				.81
4	I understand how to use this assessment.			**.72**			.65
6	I am knowledgeable about the assessment procedures.			**.81**			.63
23	I would need consultative support to implement this assessment.	-.34	.33	**-.59**			.77
24	I understand the procedures of this assessment.			**.84**			.81
28	I would require additional professional development in order to implement this assessment.		.40	-.58			.73

1	This assessment is an effective choice for understanding a variety of problems.			**-.70**		.50
7	The assessment is a fair way to evaluate the child's behavior problem.			**-.79**		.69
9	I would not be interested in implementing this assessment.			**.66**		.45
11	I would have positive attitudes about implementing this assessment.			**-.62**	.33	.73
12	This is a good way to assess the child's behavior problem.			**-.87**		.78
17	I would implement this assessment with a good deal of enthusiasm.			**-.67**		.67
21	I would be committed to carrying out this assessment.	.39		**-.64**		.75
22	The assessment procedures easily fit within my current practices.	.31		-.36	.33	.64
14	Use of this assessment would be consistent with the mission of my school.				.66	.58
13	Preparation of materials needed for this assessment would be minimal.				.71	.47
16	Material resources needed for this assessment are reasonable.				.65	.60
10	My administrator would be supportive of my use of this assessment.				.34	.19
20	Use of this assessment would not be disruptive to students.				.33	.16
25	My work environment is conducive to implementation of an assessment like this one.	.31			.29	.43

Psychometric data in support of the ITRF is promising, with published research supporting the internal consistency, temporal stability, and concurrent validity (Daniels, Volpe, Briesch, & Fabiano, 2014), as well as the classification accuracy (Daniels, Volpe, Fabiano, & Briesch, 2016) of the measure. The ITRF-G was translated from the English language ITRF according to the Kidscreen translation guidelines (see Authors, accepted with minor revisions, for a full description of the translation procedures). Initial studies of the ITRF-G focused on the psychometric properties of the short version in gate one and supported internal consistency, and classification accuracy to some degree, as well as measurement invariance across samples from the US and Germany (Authors, under review; Authors, accepted with minor revisions).

Strengths and Difficulties Questionnaire - Teacher version (SDQ-T; Goodman, 1997). The SDQ-T is a behavioral screening measure that was designed to identify those students aged 4-16 who are struggling with emotional and behavioral difficulties. Teachers are asked to rate all students in their classrooms across 25 items using a 3-point scale (i.e. not true, somewhat true, certainly true). Responses to these items are then used to generate five scale scores (i.e. Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, Prosocial Behavior), as well as a Total Difficulties score.

To date, the SDQ has been translated into over 80 languages, and the German language version was used within the current study. Psychometric studies of the German version conducted to date have supported both the construct validity and internal consistency of the measure (Bettge, Ravens-Sieberer, Wietzker, & Hölling, 2002; Saile, 2007). Furthermore, evidence of concurrent validity has been demonstrated through high correlations with selected scales of the Child Behavior Checklist - Teacher Report Form (CBCL-TRF; Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004).

Teacher Assessment Schedule for Social and Learning Behavior (LSL). The LSL (orig.: *Lehrereinschätzliste für Sozial-und Lernverhalten*; (Petermann & Petermann, 2013) is a 50-item screening measure designed to assess both the social and learning behaviors of students. Teachers are asked to indicate how frequently a student has exhibited a particular behavior using a 4-point scale (i.e. 0 = never, 3 = often). Scores are then summed in order to create 10 5-item subscales. Subscale scores falling below the 10^{th} percentile suggest the presence of a significant behavioral problem, whereas scores falling between the 10^{th} and 20^{th} percentiles suggest that the student may be at-risk for behavioral problems. Previous psychometric research has found strong evidence for the internal consistency of the measure (i.e. α=0.82 to α=.95; Gienger, 2007). Analyses conducted within the current study utilized the overall composite score.

Usage Rating Profile-Assessment (URP-A; Chafouleas, Miller, Briesch, Neugebauer, & Riley-Tillman, 2012). The URP-A is a 28-item measure designed to assess individuals' perceptions of the usability of assessment procedures. Respondents are asked to indicate the degree to which they agree with a number of statements using a 6-point Likert scale (i.e. 1 = strong disagreement, 6 = strong agreement). Responses are then used to generate six scale scores: Acceptability, Understanding, Home-School Collaboration, Feasibility, System Climate, and System Support. As noted previously, the URP-A was created by re-wording existing items designed to assess the usability

of intervention technologies to reflect the usability of assessment measures. Initial evidence in support of the URP-A supported both the six-factor structure and internal consistency of the measure; however, the reliabilities of the System Climate (α = .71) and System Support (α = .63) were found to be notably lower than the other four scales (Range = .80-.90) (Miller et al., 2012).

The URP-A was translated into the German language using a team-based four step procedure in order to ensure functional and operational equivalence of the questionnaire (Hambleton & Li, 2005). First, professional translators who were highly experienced in the translation of educational measurement tools constructed a preliminary version of the German language URP-A. Second, based on the first forward translation, a research working group consisting of the authors of the current study developed one single version by harmonizing and revising items. Third, we applied expert interviews with two primary school teachers in order to identify comprehensibility and acceptability. Fourth, we revised and modified the items based on the expert interviews, which resulted in the final German version of the URP-A.

Results

Exploratory Factor Analysis

An initial analysis of the dataset identified a total of 38 instances of missing data (1.3% of the total possible responses), which were neither restricted to particular items nor respondents. Given that the data were considered to be missing at random, the decision was made to impute missing values using multiple imputation (Enders, 2001). A total of 10 datasets were generated and the resultant values were combined in order to produce each imputed estimate. Next, the data were examined to ensure that they were appropriate for conducting a factor analysis. First, the correlation matrix was examined for either signs of multicollinearity (i.e. inter-item correlations above .80) or low communalities (i.e. inter-item correlations above .30 with fewer than three items). No items were found to be problematic with regard to either criterion. Second, the anti-image correlation matrix was examined to ensure that the measure of sampling adequacy (MSA) for all items was above .60 (Pett et al., 2003). The MSA represents the degree to which the item is correlated with other items in the measure, and no problematic items were identified. Third, both the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (.82) and Bartlett's Test of Sphericity (χ^2 (378) = 2017.71, $p < .001$) suggested that (a) there were no problems with the size of the sample and (b) the matrix was factorable.

Exploratory factor analysis was conducted using SPSS 23.0 using principal axis factoring with an oblique rotation, given that the factors were expected to be correlated with one another. Decisions regarding the number of factors to extract were made by considering multiple factors. Examination of the scree plot seemed to suggest an elbow in the data between the fifth and sixth factors. In addition, eigenvalues at or above 1.0 were identified for five factors and the results of parallel analysis suggested a five-factor solution. Given that extraction of six factors was found to result in a one-item factor; the decision was ultimately made to extract five factors. These five extracted factors accounted for 60.68% of the common variance in items.

Subsequent to factor extraction, indicators were considered in order to identify any potentially problematic items. First, the pattern coefficient matrix was reviewed to identify any items that either (a) loaded poorly on the primary factor (i.e. pattern coefficient < .45) or (b) demonstrated strong factor loadings on more than one factor. This resulted in the removal of Items 10, 20, 22, and 25. Second, the final item communalities were reviewed in order to identify any items for which the proportion of item variance accounted for by the extracted factors was found to be substantially low. No additional items were deleted at this stage.

Reliability Estimates

Reliability analyses were next conducted for each of the five extracted factors. First, the inter-item correlation matrices were examined in order to identify any items that were either minimally correlated with other items in the scale or which demonstrated notably high correlations with other items in the scale. Given the high correlation between Items 23 and 28 (r = .86) within Factor III, the decision was made to delete Item 28 from the final scale. Next, we looked to determine whether the deletion of any individual items would result in a significant improvement in scale reliability; however, no items were found to be problematic.

Acceptable levels of reliability were found for all subscales of the German URP-A (see Table 3). Subscale I (α = .88) was comprised mostly of items from the Feasibility subscale of the URP-A, designed to assess the degree to which potential barriers to implementation of an assessment may be present (e.g., requires too much time, is too complex). However, two additional items were found to load on this factor, which were previously considered to measure System Support (i.e., *I would need additional resources to carry out this assessment*) and System Climate (*These assessment procedures are consistent with the way things are done in my system*). The mean score for this subscale was 3.97, suggesting that teachers found the screening measures to be somewhat feasible to implement.

Table 3. ***Reliability Estimates for the German URP-A***

Subscale	Items	α	95% CI (α)	Subscale Mean
Feasibility	2, 3, 8, 18, 19, 26	.88	.83, .91	3.98
Home-School Collaboration	5, 15, 27	.88	.83, .91	5.29
Understanding	4, 6, 24, 28	.90	.86, .93	3.24
Acceptability	1, 7, 9, 11, 12, 17, 21	.90	.87, .93	3.97
Omitted Factor[a]	13, 14, 16	.75	.66, .83	4.52

Note. [a] This factor was omitted from the final measure due to lack of conceptual similarity among some items, as well as the presence of a potential wording artifact.

Subscale II (α = .88) was found to be consistent with previous psychometric evaluations of the URP-A. Within this Home-School Collaboration subscale, the

three items were designed to assess the degree to which support and collaboration from families is needed in order to implement the assessment procedures. The mean score for this subscale was 3.24, indicating that teachers slightly disagreed that support from families would be needed in order to implement the screening assessments.

Subscale III ($\alpha = .90$) was comprised mostly of items from the Understanding subscale of the URP-A. These items were designed to assess the degree to which a potential user understands how to implement an assessment procedure and feels that she can do so independently. However, two additional items were found to load on this factor, which were previously considered to measure System Support (i.e., *I would need consultative support to implement this assessment, I would require additional professional development in order to implement this assessment*). The mean score for this subscale was 5.29, suggesting that teachers felt that they had sufficient knowledge in order to implement the screening assessments.

Subscale IV ($\alpha = .90$) was found to be consistent with previous psychometric evaluations of the URP-A. Within this Acceptability subscale, items were designed to assess the degree to which potential users perceive the assessment to be an appropriate assessment tool and see it as something that they are interested in using themselves. The mean score for this subscale was 3.98, indicating that teachers would be somewhat interested and willing to implement the screening assessments described.

The lowest level of reliability was found for Subscale V ($\alpha = .75$), which was comprised of three items from both the Feasibility and System Climate subscales of the URP-A (i.e. *Preparation of materials needed for this assessment would be minimal*, *Use of this assessment would be consistent with the mission of my school*, *Material resources needed for this assessment are reasonable*). Given the lack of conceptual similarity among some items, as well as the presence of a potential wording artifact (i.e. focus on materials), the decision was made to omit the fifth factor from the final measure.

Discussion

The URP-A was designed to assess consumers' perceptions of the usability of those assessment procedures that may be employed in school settings. The English language version of the measure consists of 28 items designed to assess a total of six subscales: Acceptability, Understanding, Feasibility, Home-School Collaboration, System Support, and System Climate. Within the current validation of the German version, eight of these items were removed, resulting in a 20-item measure assessing the four primary factors of Acceptability, Understanding, Feasibility, and Home-School Collaboration. All four subscales were found to demonstrate as high—if not higher—levels of reliability than were found for the original URP-A, suggesting great promise for applied use in German schools.

The two subscales that were found to be the most consistent across the initial (i.e., Miller et al., 2013) and current evaluations were those of Home-School Collaboration and Acceptability. Without changing any of the three items within the Home-School Collaboration subscale, a notable improvement in reliability was evidenced (i.e. from .83 to .88) within the current sample. These results are promising from a psychometric standpoint; however, it is also worth noting that the mean score for this subscale was found to be fairly low. That is, respondents tended to slightly dis-

agree that support from families would be needed in order to carry out the described assessments. This finding does not come as a great surprise, given that only teacher ratings were used within the ITRF, SDQ, and LSL. If, however, parent ratings were gathered as part of the screening process (as is possible within the SDQ, for example), results for this subscale may be more informative. Within the Acceptability subscale, two of the items were deleted due to low factor loadings (*Use of this assessment would not be disruptive to students, The assessment procedures easily fit in with my current practices*); however, the resultant 7-item subscale was found to demonstrate the same strong level of reliability as the original 9 items.

In contrast, a greater number of changes in item content were found for the remaining subscales. Most notable of these changes was the fact that the two factors designed to assess system-level considerations did not emerge as distinct factors. Whereas some of the items originally belonging to the System Climate and System Support subscales were deleted from the measure entirely, other items were found to load more strongly on either the Feasibility or Understanding subscales within the current evaluation. First, one of the items originally designed to assess the degree to which an assessment is believed to be compatible with the culture of the school (i.e. System Climate) was found to relate more strongly to those items assessing Feasibility (i.e., *These assessment procedures are consistent with the way things are done in my system*). Because the content of this item was believed to be conceptually consistent with what the Feasibility subscale was designed to measure (i.e., the degree to which consumers believe that the assessment requires a reasonable amount of time, personnel, and/or resources to implement), the item was therefore retained. In addition, two items originally designed to measure System Climate were deleted due to low factor loadings (*My administrator would be supportive of my use of this assessment*, *My work environment is conducive to implementation of an assessment like this one*) and one was deleted after loading on the fifth factor that was ultimately omitted (*Use of this assessment would be consistent with the mission of my school)*. The exact reasons why these three items did not perform as well as within the current evaluation are unknown; however, one possible explanation may be that teachers may have been thinking more about how the tool would fit within their own classroom than how it might be incorporated within the larger school setting, given the analog nature of the task.

The System Support subscale was designed to assess the degree to which external supports are believed to be necessary in order to use an assessment. Similar to the System Climate factor, two of the original items within this subscale were found to load more strongly on other factors. The item *I would need additional resources to carry out this assessment* was therefore moved to the Feasibility subscale and the item *I would need consultative support to implement this assessment* was moved to the Understanding subscale. Again, these changes were believed to maintain conceptual consistence with the intended focus of the scales and therefore not seen as problematic. The one additional item that had originally loaded on the System Support subscale (*I would require additional professional development in order to implement this assessment*) was deleted from the measure when high inter-item correlations suggested it was redundant with the item assessing the need for consultative support.

Limitations and Directions for Future Research

Although the results of the current study suggest great promise for the use of the URP-A within German school settings, there are a few limitations of the study that should be noted. First, the size of the sample may be considered somewhat small given the number of items within the measure. Guidelines for conducting an EFA tend to specify that there should be 5 respondents in the sample for every item in the measure (Fabrigar, Wegener, MacCallum, & Strahan, 1999), which would equate to 140 respondents for the 28-item URP-A. The fact that the sample utilized in the current study was somewhat smaller (i.e. 101) than outlined by these recommendations suggests the need for further validation with a larger sample in order to ensure the replicability of results. It is worth noting, however, that no problems with sample size were suggested by the Kaiser-Meyer-Olkin Measure of Sampling Adequacy, which was found to be sufficient given the number of items.

Second, as noted previously, the current investigation was part of a larger study in which teachers were provided with vignettes describing both the ITRF and a second screening behavioral measure in order to examine the perceived usability of each tool. Although only one set of URP-A responses was used for each teacher within this validation study, it is possible that those ratings were influenced by previous ratings to some degree. For example, teachers may have rated the second screening measure more stringently relative to the first or may have become less focused on the assessment task by the time that they read the second vignette. Such potential issues would be possible to avoid if each teacher only received a single, randomly-assigned vignette upon which to base her ratings.

Third, although multiple vignettes were used in order to ensure sufficient variability in responses, all vignettes were focused on behavioral screening measures. As such, it is possible that different results may be found if teachers were asked their perceptions of a wider range of assessment tools, such as those used to assess academic domains or those used for diagnostic purposes. Additional investigation is therefore warranted in order to ensure that the underlying factor structure holds consistent across the full range of school-based assessment tools.

Finally, the psychometric information generated through the current investigation was limited to evidence of construct validity and internal consistency reliability. As such, it is important for future research to explore additional aspects of psychometric adequacy including the test-retest reliability and construct validity of the data generated through the URP-A.

Implications for Practice

The URP-A was developed to serve two key purposes in understanding and promoting usage of school-based assessment tools. First, within a research context, the URP-A was designed to be assessment-neutral, in that it can be completed with regard to a range of different assessment tools. Over the years, many researchers have aimed to include an assessment of social validity within their studies, so as to illustrate the degree to which the assessment or intervention was not only effective but also acceptable to the intended user. Unfortunately, however, these social validity measures are often researcher-created and therefore inconsistent from one investigation to the next. Use of a standardized tool such as the URP-A therefore allows for

more direct comparisons to be made across individual research studies in order to understand the way in which tools are perceived relative to one another.

Second, within applied settings, the URP-A may be used to assist consultants in the efficient development of sustainable assessment plans. Typically, when a consultant meets with a teacher to devise an assessment plan, the teacher is asked about her perceptions of the plan in a more global, holistic way (e.g., "What do you think about the plan?"). As such, it is not always possible to identify potential barriers to implementation prior to actually putting the plan into action. However, by having the classroom teacher complete the URP-A early on in the process of plan development, it is possible to pinpoint what factors—either individually or in combination—may prevent future success. Knowing early on, for example, that a teacher does not fully understand what is being asked of her, or that she perceives an assessment tool to require too large of a time commitment, would allow the consultant to proactively make necessary and appropriate changes to the assessment plan in order to enhance the probability of effective usage.

References

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

Broughton, S. F., & Hester, J. R. (1993). Effects of administrative and community support on teacher acceptance of classroom interventions. *Journal of Educational & Psychological Consultation, 4*, 169-177.

Buston, K., Wight, D., Hart, G., & Scott, S. (2002). Implementation of a teacher-delivered sex education program: Obstacles and facilitating factors. *Health Education Research, 17,* 59-72.

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

Daniels, B., Volpe, R. J., Briesch, A. M., & Fabiano, G. A. (2014). Development of a problem-focused behavioral screener linked to evidence-based intervention. *School Psychology Quarterly, 29*, 438-451.

Daniels, B., Volpe, R. J., Fabiano, G. A., & Briesch, A. M. (2016). Classification Accuracy and Acceptability of the Integrated Screening and Intervention System Teacher Rating Form. *School Psychology Quarterly*. http://doi.org/10.1037/spq0000147

Eagle, J. W., Dowd-Eagle, S. E., Snyder, A., & Holtzman, E. G. (2015). Implementing a multi-tiered system of support (MTSS): Collaboration between school psychologists and administrators to promote systems-level change. *Journal of Educational and Psychological Consultation, 25*, 160-177.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*, 128-141.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities: Research & Practice, 18*, 172-186.

Gienger, C. (2007). Lehrereinschätzliste für Sozial- und Lernverhalten (LSL).: Göttingen: Hogrefe. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *55*, 209-210.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581-586.

Grosche, M. & Volpe, R. J. (2013). Response-to-intervention (RTI) as a model to facilitate inclusion for students with learning and behaviour problems. *European Journal of Special Needs Education, 28*, 254-269.

Kazdin, A. E. (1980). Acceptability of alternative treatments for deviant child behavior. *Journal of Applied Behavior Analysis, 13*, 259-273.

Logan, P., & Skinner, C. H. (1998). Improving students' perceptions of a mathematics assignment by increasing problem completion rates: Is problem completion a reinforcing event? *School Psychology Quarterly, 13*, 322-331.

Mautone, J. A., DuPaul, G. J., Jitendra, A. K., Tresco, K. E., Junod, R. V., & Volpe, R. J. (2009). The relationship between treatment integrity and acceptability of reading interventions for children with Attention-Deficit/Hyperactivity Disorder. *Psychology in the Schools, 46*, 919-931. doi: 10.1002/pits.20434

Miller, F. G., Neugebauer, S. R., Chafouleas, S. M., Briesch, A. M., & Riley-Tillman, T. C. (2013). *Examining innovation usage: Construct validation of the Usage Rating Profile - Assessment.* Poster presentation at the American Psychological Association Annual Convention, Honolulu, HI.

Petermann, -U., & Petermann, -F. (2013). *Lehrereinschätzliste für Sozial- und Lernverhalten* (2nd ed.). Retrieved from http://search.ebscohost.com/login.aspx? direct=true&db=pdx&AN=PT9005584&site=ehost-live

Power, T. J., Hess, L. E., & Bennett, D. S. (1995). The acceptability of interventions for Attenton-Deficit Hyperactivity Disorder among elementary and middle school teachers. *Journal of Developmental and Behavioral Pediatrics, 16*, 238-243.

Reimers, T. M., Wacker, D. P., & Koeppl, G. (1987). Acceptability of behavioral treatments: A review of the literature. *School Psychology Review, 16*, 212-227.

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

Witt, J. C. (1986). Teachers' resistance to the use of school-based interventions. *Journal of School Psychology, 24*, 37-44.

Witt, J. C., Noell, G. H., LaFleur, L. H., & Mortenson, B. P. (1997). Teacher use of interventions in general education settings: Measurement and analysis of the independent variable. *Journal of Applied Behavior Analysis, 30*, 693-696.