Received: 13 October 2018 | Revised: 2 January 2019 | Accepted: 4 January 2019

DOI: 10.1002/mpr.1771

ORIGINAL ARTICLE

WILEY

Examining item content validity using property fitting analysis via multidimensional scaling

Cody Ding[1,2]

[1] Shenyang Normal University, College of Education, Shenyang, China

[2] Department of Education Sciences & Professional Programs, University of Missouri–St. Louis, St. Louis, Missouri

Correspondence
Cody Ding, Education Sciences and Professional Programs, University of Missouri–St. Louis, St. Louis, MO.
Email: dingc@umsl.edu

Abstract

Objectives: Multiple-item measuring instruments are frequently used in a wide range of disciplines for the purpose of research in substantive areas. The quality of items in these instruments determine to a large extent whether the results are trustworthy. In this paper, we suggested to use property fitting analysis to evaluate the appropriateness of items content validity based on explicit item property criteria.

Methods: Using Center for Epidemiologic Studies Depression scale as an example, item property fitting analyses via multidimensional scaling model was used to quantitatively evaluate the properties of items based on rating data from 12 counselors.

Results: The results of the analyses indicated that using explicit item property criteria to select items for subsequent psychometric analyses improved the item quality in terms of reliability and factor structure.

Conclusions: Item property fitting analysis seemed to provide the researcher a viable quantitative method when evaluating item content validity.

KEYWORDS

explicit item property criteria, item content validity, item property fitting analysis, multidimensional scaling

1 | INTRODUCTION

Researchers in a wide range of disciplines are frequently involved in the development and revision of multiple-item measuring instruments such as scales, tests, inventories, questionnaires, surveys, subscales, or testlets. Scores obtained from these measuring instruments are usually employed in subsequent analyses that address substantive research questions. To a considerable degree, the quality of items in these instruments determines the extent to which the analyses and modeling efforts produce trustworthy results. In a sense, measuring instruments used in social sciences research are analogous to diagnostic equipment (e.g., X-ray or CT scanner) used in medical settings, in which accurate and reliable results depend on the quality of the diagnostic instruments used.

In order to construct scales or questionnaires with items of high psychometric quality, researchers must engage in many activities aimed at building initial versions of the instrument items and then repeatedly revise them to improve their performance (e.g., Kanter, Mulick, Busch, Berlin, & Martell, 2007; Lehmann-Willenbrock & Kauffeld, 2010; Lu & Gilmour, 2006; Murray, Booth, McKenzie, & Kuenssberg, 2015). These psychometric activities may be classified into several categories: (a) activities involved in development of initial item pool based on content validity (e.g., Beauchamp et al., 2010; Kessler, Andrews, Mroczek, Ustun, & Wittchen, 2006; Lu & Gilmour, 2006), (b) analysis of factor structure of items with respect to construct validity, including convergent and divergent validity, item biases, and measurement invariance (e.g., Ferrer, Balluerka, & Widaman, 2008; Hughes, Betka, & Longarzo, 2018; Storch, Rasmussen, Price, Larson, & Murphy, 2010; Walker, 2010; Williams & Polaha, 2014), (c) analysis of criterion-related validity (e.g., Klassen et al., 2009), and (d) analysis of reliability (e.g., Funk, Huebner, & Valois, 2006). Among these psychometric activities, most of them use quantitative methods to examine item behavior or property in the analysis. Analytical methods most seen in the research literature include factor analysis (both exploratory and confirmatory analysis) for construct validity (e.g., Beauchamp et al., 2010; Lenz, Balkin, Gómez Soler, & Martínez, 2015), correlation analysis for convergent or

divergent validity (e.g., Burns & Rapee, 2016), and regression analysis for criterion-related validity (e.g., Burns & Rapee, 2016). In addition, analysis based on item response theory is also used to examine item property such as item trait, item discrimination, or differential item functioning in some cases (e.g., Draheim, Harrison, Embretson, & Engle, 2017, March 9; Murray et al., 2015).

All these psychometric analyses are conducted after the initial item pool is developed (e.g., Chao & Green, 2011). Typically, the initial item pool is compiled based on the idea of content validity, which states that the items used should be representative of a specific concept domain. For example, items of a depression measure should ideally represent the behavioral manifestation of depression (however, it is conceptualized). In the psychometric literature, development of initial items usually does not use quantitative methods for selecting and examining the items with respect to the content validity or what the items can assess. For example, for assessing the methodological quality of studies on measurement properties of instruments, Mokkink et al. developed the COSMIN checklist (2010). The COSMIN checklist contains standards for evaluating instrument properties such as construct validity or reliability. All these standards are accompanied by statistical methods for analysis except for content validity. That is, although there are standards for evaluating content validity, no statistical methods are discussed for evaluating item properties with respect to content validity. In the measurement literature, content validity is examined, instead, by a series of qualitative evaluation steps. Although there may be some variations to these steps, the process often includes the following: First, initial items are compiled or developed based on existing literature on a specific area (e.g., stigma about mental health) or adoption of existing items in that area. Sometimes, focus groups are used to elicit items. Second, an expert panel is used to review, evaluate, and revise the items in the item pool based on experts' perception of what items assess. In this step, researchers sometimes use quantitative methods (e.g., content validity scores) to quantify the content validity (Zamanzadeh, Ghahramanian, Rassouli, Abbaszadeh, & Nikanfar, 2015). Finally, a focus group of stakeholders is interviewed with respect to the appropriateness of wording or misunderstanding of what items might assess (i.e., clarity and relevancy). This combined process produces initial item pool for subsequent psychometric analyses. Perusing articles on the development of a new instrument in *Psychological Assessment*, for example, shows that the development of initial item pool is done through some combination of these activities (e.g., Burns & Rapee, 2016; Chao & Green, 2011; Williams & Polaha, 2014).

One of the issues in this process is that there is no single set of explicit standards or criteria of item property researchers use to evaluate what items may assess with respect to content validity, particularly in the process of expert review. Experts are often asked to review the items based on their knowledge of the area, but a researcher may have no idea of what standards or criteria of item property the experts are using to judge the appropriateness of the item. The purpose of this study was to propose a process of setting explicit criteria of item property for experts or any individuals involved in appraising the items to use when evaluating the item with respect to what each item should assess. Then we can quantify the results so that we can visualize how the items function with regard to these explicit criteria of item properties. Given that the initial item development plays an important role for content validity as well as laying a foundation for subsequent psychometric analyses (Beck, 1999; Zamanzadeh et al., 2015), it is necessary to stipulate explicit criteria of item property when evaluating what items can assess.

2 | ESTABLISHING EXPLICIT ITEM EVALUATION CRITERIA

When an expert panel is used to review, evaluate, and revise a pool of initial items, it is important to be explicit about what we want them to focus on, that is, what criteria of item property we want them to use to review and evaluate the appropriateness, suitability, relevancy, or necessity of items in measuring a specific construct. Terwee et al. (2007) call this quality criteria. What is often lacking when evaluating items in this phase is explicit criteria of item properties for which a set of good items should possess. For assessment of content validity, it is often recommended that a clear description be provided on measurement aim, the concepts that are being measured, and the item properties (e.g., Schellingerhout, Verhagen, Heymans, Koes, & de Vet, 2012). However, no explicit criteria of item property have been stipulated as to how items should be judged in this process. Instead, a holistic approach (i.e., based on expert knowledge) of appraising the item properties is typically used. Thus, as part of an effort to improve content validity, we should develop an explicit set of criteria of item properties for evaluating the items before we move forward to subsequent psychometric analyses such as reliability or various validities.

When establishing explicit item property criteria for evaluating items with respect to content validity, we must determine which item properties should be included in the criteria. Although such determination depends on the specific content area (e.g., adolescent health, anxiety, or anger), the general explicit criteria can focus on the following item properties based on suggestion by Mokkink et al. (2010).

a. Whether items are relevant for the purpose of the instrument (e.g., diagnostic, discriminative, evaluative, or predictive).

The aim of measurement instruments is important because different items may be valid for different purposes. For example, items that may be good for the purpose of diagnosis may not be valid for the purpose of prediction. For assessment of mental health, many instruments are designed for diagnosing the presence or absence of certain mental health problems. However, items comprising these instruments may not be suitable for predicting the occurrence of future mental health problems. Therefore, we should be explicit about the goal of instruments so that items with that property can be developed accordingly. For example, the well-known Rosenberg Self-Esteem Scale (Rosenberg, 1965) is widely used for assessing feelings about the self. However, it is unclear what the purpose of this 10-item instrument is. Is it currently used for diagnosing various levels of self-esteem, for differentiating people with high self-esteem from those with low self-esteem, for screening individuals for depression, or for predicting one's self-esteem? Although it is possible that the

instrument can accomplish all these purposes, it is necessary to clearly understand item properties of the instrument with respect to diagnosis, evaluation, or prediction.

b. Whether items are relevant for the study population (e.g., age, gender, and symptom characteristics).

It is necessary to evaluate the relevance and applicability of the instrument items that are specific to certain groups of individuals. For example, Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977) is designed for screening depressed mood in the general population but items may not be suitable to adolescents. Thus, items are required to be adapted for use with adolescent populations. The CES-D for Children is designed for such a purpose (Faulstich, Carey, Ruggiero, Enyart, & Gresham, 1986). When we develop an item pool, we must consider age group, gender, or symptom characteristics to determine which item properties are most appropriate.

On the basis of these general explicit criteria of item property for examining items, we can construct a matrix of item property that can be used by expert panels when judging the properties of the items in the instrument. The matrix of item properties is made of *k* items by *m* properties, as shown below:

	Property_1	Property_2	...	Property_m
item_1				
item_2				
...				
item_k				

The matrix explicitly articulates item properties the panel of experts should focus on. A rating scale can be employed to quantify the degree to which each item matches these explicit criteria of item properties, such as using content validity index (Davis, 1992). Thus, each expert in the panel provides a rating for each item with respect to a particular property. The average of each item's ratings across experts can be used as data input for a subsequent analysis of these ratings. The results of analysis can then be used to guide the selection of items for further psychometric analyses.

Using the example of the CES-D scale (Radloff, 1977) should make these points clearer. CES-D is a 20-item scale that is designed to assess depressive symptomology in the general population (Radloff, 1977). The items are selected from a pool of items from previous depression scales, and it covers the following areas: depressed mood, feelings of guilt and worthlessness, feelings of helplessness and hopelessness, psychomotor retardation, loss of appetite, and sleep disturbance. It includes four positively worded items. Radloff (1977) indicated that the scale is useful for epidemiologic studies of depression.

For this set of 20 CES-D items, we constructed a matrix of item property based on the general criteria of good content validity, as suggested by Monkkink et al. (2010). Because the purpose of CES-D is to screen depression in the general population, we focused on seven item properties we believed to be relevant and necessary for these items to possess. These properties are typicality (how typical this symptom is for depression), frequency (how often this symptom occurs in depression), differentiating (can this symptom differentiate individuals with depression from those without it), sensitivity (is this symptom sensitive in assessing depression), specificity (is this symptom specific to depression), fit (is this symptom appropriate to use in the general population), and predictive (can this symptom predict depression). Of course, there are some alternative criteria of item property that can be used. The seven item properties used here are mainly for didactic purposes. The key point is that explicit criteria of item property are critical for experts or individuals to use when judging the relevance and importance of items to be included in the final item pool.

3 | QUANTIFYING THE ITEM PROPERTY

To analyze the item property using data obtained from the matrix of item property, three types of analyses can be conducted. First, we can conduct the analysis of interrater reliability (IRR) or agreement to examine the degree of similarity in rating each property across items by a group of raters. This is a common practice in assessing judgment consistency across raters. In our current example of rating item property using 20 CES-D items, this analysis results in seven indices of interrater agreement, one for each item property. The rating data are ordinal or interval in nature; therefore, the intraclass correlation (ICC) coefficient can be calculated as a measure of IRR reflecting the accuracy of the rating process, as suggested by Stolarova, Wolf, Rinker, and Brielmann (2014) and Hallgren (2012). Due to the seven ICCs, confidence intervals (CIs) for all ICCs are calculated in order to assess whether they differ from each other.

Second, we can assess how raters perceive or interpret these item properties; that is, do raters perceive or interpret these item properties differently? This aspect of analysis is interesting because it can help us examine any systemic individual differences in using these properties to evaluate the item quality. Ideally, we would like to see all raters use these item properties in the same way, indicating minimum rater biases. We consider this assessment an analysis of rater bias.

Third, after we examine the IRR and rater bias, which serve as preliminary steps ensuring the accuracy of the rating process, we can conduct item property analysis. In this type of analysis, we examine the degree to which each item corresponds to these item properties. A close correspondence between items and their properties indicates good content validity in regard to these properties. In the following, we demonstrated these analyses using rating data of CES-D items based on the matrix of item property.

4 | METHODS

4.1 | Participants

For the purpose of judging properties of items, we recruited 12 counselors working in various mental health agencies (e.g., hospital or private clinics) in the Midwest region of the United States. All

participants were female, with average age being 34.8 (*SD* = 3.6). Among them, four were African-Americans.

The original version of the 20-item CES-D scale was also administered to a group of 234 college students, with 65.6% being female. The mean age of the participants was 20.5 years old (*SD* = 3.56). In this sample, 25.4% were freshmen, 26.2% were sophomores, 14.3% were juniors, and 33.2% were seniors. We used this sample to examine the factor structure of items that were deemed to possess item properties.

4.2 | Procedure

Each counselor was asked to appraise the item quality of 20 items of CES-D scale with respect to the seven item property indicators (i.e., typicality, frequency, differentiating, sensitivity, specificity, and predictivity). The item property matrix is shown in Appendix A. Specifically, they were asked to make independent ratings of items on the CES-D scale based on the definition of each item property. A rating scale of 0 to 9 was used, with 0 indicating *not a good item* and 9 indicating *a very good item*. Participation in the study was voluntary.

4.3 | Data

Because CES-D scale is a 20-item instrument, the rating data were a 20 by 7 matrix of rating scores, one for each participant (or rater), totaling 12 such matrices. The subsequent analyses were based on these matrices.

4.4 | Analysis design

In carrying out analysis of item property, three types of analyses were performed, as mentioned above. The IRR was estimated using an ordinal metric ICC (Hallgren, 2012). ICCs incorporated the magnitude of the disagreement in computing the IRR estimates. Thus, IRR was assessed using a two-way mixed, absolute, average-measures ICC (McGraw & Wong, 1996). IRR was separately calculated across items for each property as an estimate for the accuracy of the rating process. Higher ICC values indicated greater IRR, with an ICC estimate of 1 indicating perfect agreement.

The analysis of rater bias and of item property was based on multidimensional scaling model (MDS). The detailed discussion of this method is beyond the scope of this paper, and we only discussed some key points that were more relevant to the current topic. Readers who are interested in the method were referred to the book by Ding (2018). For the analysis of rater bias, that is, whether counselors perceived (or interpreted) these item properties similarly or differently, we used individual differences scaling (INDSCAL) within MDS. As indicated by Horan (1969) and Caroll and Chang (1970), we can presume that each individual or a group of individuals may perceive an item property (e.g., typicality) differently based on understanding of various attributes of typicality. We have a latent or group space (i.e., latent common space) that consists of all the attributes the individuals happen to use. Each individual's space can now be thought as a special case or subcase of the group space because the individual is using some or part of the total available attributes of typicality. Each individual's space can be termed as his or her private space.

To operationalize this idea of individual differences with respect to potential differences in interpretation of typicality into MDS analysis, we assume that each individual attaches a different weight to each dimension of typicality that represents his or her degree of salience, attention, or importance of that dimension when he or she makes rating. Thus, each individual has his or her unique set of weights. For example, an individual who attaches equal salience to each of the dimensions will have a set of weights of the equal or very similar value. In contrast, individuals who attach a different weight to each dimension systematically deviate from the group space into their own private space. Accordingly, INDSCAL is a method of modeling how individuals vary in terms of differing weights being associated with the same dimensions. The model fit can be assessed using Kruskal's Stress-1 value (Kruskal, 1964) and Dispersion Accounted For (Data Theory Scaling System Group, n.d.). Values of both indices range from 0 to 1, with 1 indicating the perfect model fit.

To quantitatively evaluate item properties, we performed property fitting analysis using the MDS model. Property fitting analysis inputs both a configuration of items and property ratings of the same set of items (i.e., these ratings are estimates of different properties of the items). In the analysis, each property is presented as a vector through the configuration of item points, indicating the direction over the space in which the item property is increasing. Thus, the relative length of the vector indicates the degree of model fit, with longer length indicating a better model fit. The model fit can also be assessed by R^2.

The analyses of IRR and rater bias were done using SPSS 25 (IBM Corp, Released, 2017) and property fitting was conducted using SAS (2013).

5 | RESULTS

5.1 | Interrater reliability

IRR was calculated for each item property as an estimate for the accuracy of the rating process. The resulting average ICC was in the excellent range, ICC_{ave} = 0.89 (Cicchetti, 1994), indicating that these counselors had high levels of agreement and suggesting that rating of items across all item properties was very similar across raters. The CI of ICCs for each item property is shown in Figure 1. These CIs were overlapping, indicating that they do not differ from each other. The only exception was the ICC CI of item property “Fit”; its CI was much wider and its IRR was lower than the other ones, suggesting that rating of items using this item property was much less consistent across these counselors.

5.2 | Rater interpretation bias

A two-dimensional INDSCAL solution yielded an excellent model fit, with Stress-1 value being 0.046, and Dispersion Accounted For being 0.99. The results indicated that a two-dimensional space had a good fit to the configuration of the 20 CES-D items.

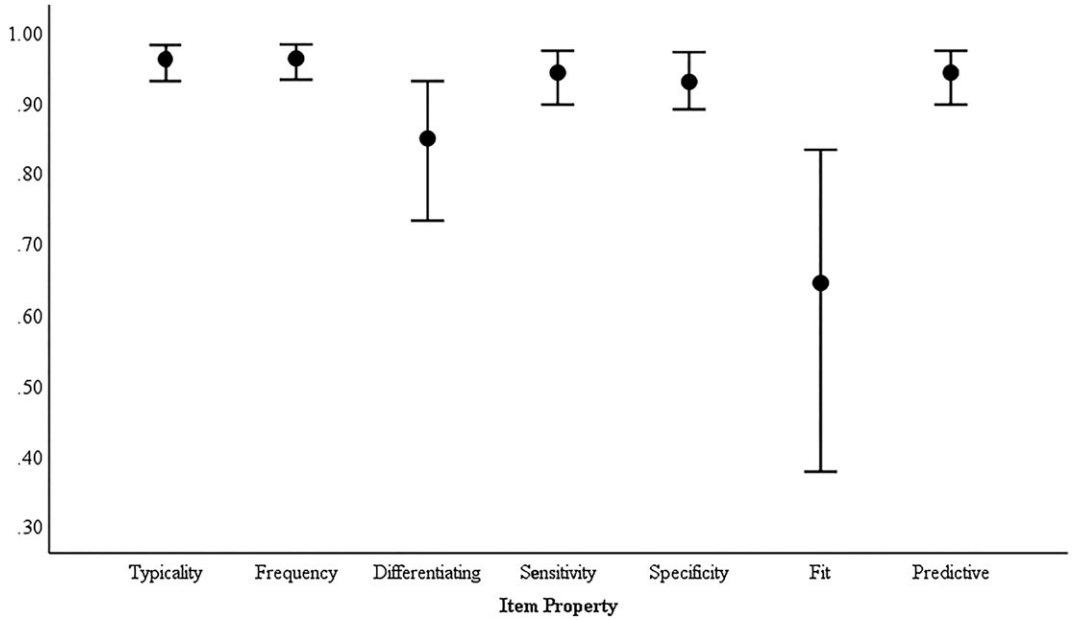

FIGURE 1 Comparison of interrater reliability. Intraclass correlation coefficients (ICCs, represented as dots) and corresponding confidence intervals at $\alpha = 0.05$ (CIs, represented as error bars) across seven item properties. Overlapping CIs indicate that the ICCs did not differ systematically from each other except for one item property

The analysis produced a visual display of how these counselors as a group perceived the item properties (i.e., group configuration), the configuration weights, and the configuration for each counselor. Figure 2 shows the group configuration of the seven item properties as well as configuration weights. As can be seen in Figure 2a, seven item properties seemed to form two clusters along two dimensions: properties focusing on the item itself and properties focusing on comparisons to others. The configuration weight in Figure 2b showed one configuration weight rather than 12 different configuration weights, and the weight was almost on a diagonal line. This finding indicated that the counselors attached equal salience in using these item properties to rate CES-D items. Therefore, no systematic bias (i.e., uniqueness) in interpretation of these item properties was found. This result coincided with the results of the high ICC, suggesting that a minimal amount of systematic measurement error was introduced by the independent raters. Accordingly, item property ratings were deemed to be suitable for use in the item property fitting analysis.

5.3 | Item property

In item property fitting analysis, item properties were linearly regressed onto configurations of CES-D items, yielding seven vectors of item properties superimposing onto the configuration of 20 items. Each item property had a good fit to the configuration of items, with R^2 ranging from 0.93 to 0.99. A visual display of results of property fitting is shown in Figure 3. Two things were noticeable, as indicated in Figure 3. First, all seven item properties were clustered together in the same direction, indicating that these item properties were perceived to be highly correlated. The vector length of item property "Fit" was shorter than that of the rest, indicating that this item property did not fit the model as well as the other item properties. Second, there seemed to be three distinct clusters of the 20 CES-D items. Items 4, 8, 12, and 16 were positively worded items, which formed one cluster. In the direction between 6 o'clock and 8 o'clock, Items 1, 2, 5, 10, 11, 13, and 15 formed a second cluster. The rest of the nine items formed a third cluster. All vectors of item property were pointing to the direction of these nine items, and this pattern indicated the direction over the space in which the property was increasing. Thus, all seven properties seemed to be judged as more descriptive of these nine items.

Obviously, four positively worded items were facing the opposite direction of the item property vectors, indicating that these item properties did not match items used to represent depression symptoms. The other seven items lay between these two clusters of items, indicating a somewhat poor fit of item properties to these items; that is, these items did not describe depression well in the context of these seven properties.

5.4 | Factor structure and reliability

For the nine items matching the item properties, we may consider whether or not they demonstrated the expected reliability and factor structure. To investigate this question, we performed an exploratory factor analysis, and the result suggested a clean one-factor structure. Figure 4 shows the factor structure and the factor loading. The reliability estimate of scores from these nine items was 0.88. In contrast, the reliability estimate of scores from the 20-item instrument was 0.82. This suggested that improved item quality based on the explicit criteria of item selection resulted in a better instrument in terms of reliability.

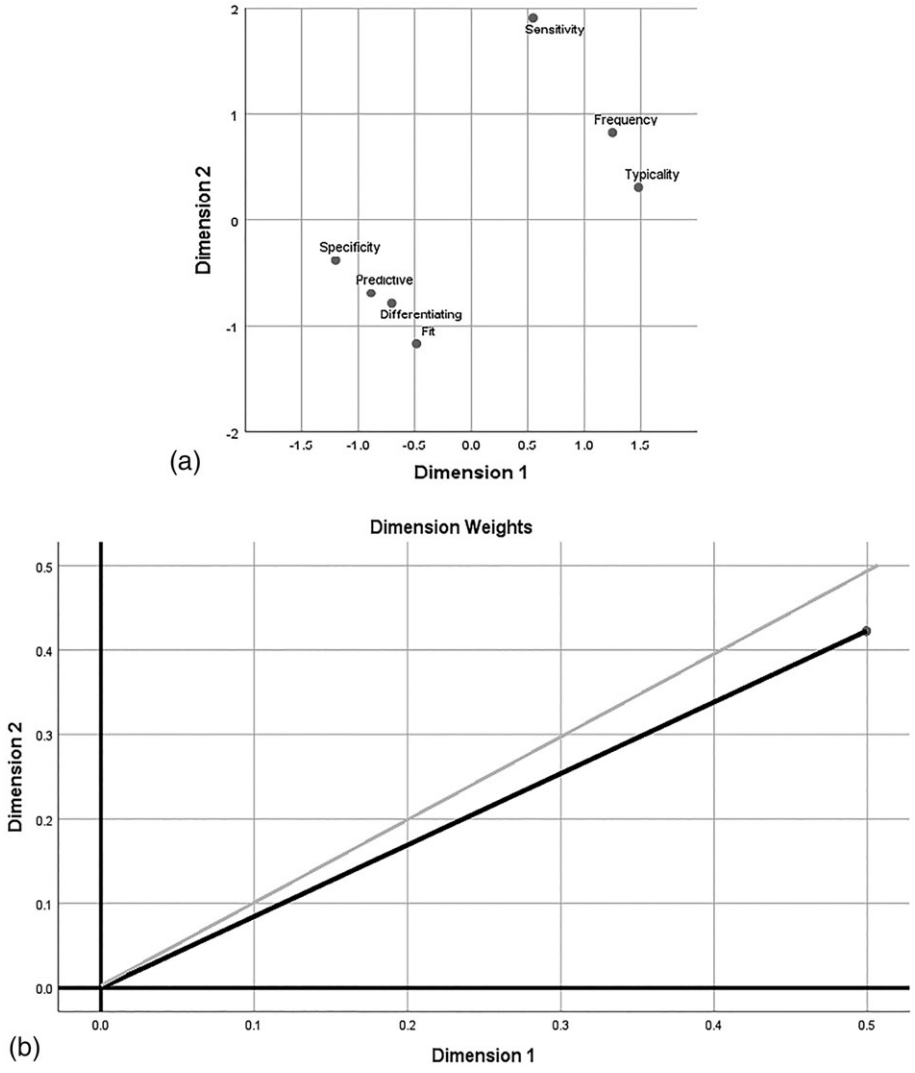

FIGURE 2 Group configuration and configuration weight from the INDSCAL analysis. Configuration weight along the diagonal suggests equal salience or focus in using item properties to rate Center for Epidemiologic Studies Depression items. The gray line indicates the expected equal salience along the two dimensions. The great departure from this theoretically equal presentation indicates the individual differences (or biases). (a) Group configuration and (b) configuration weights

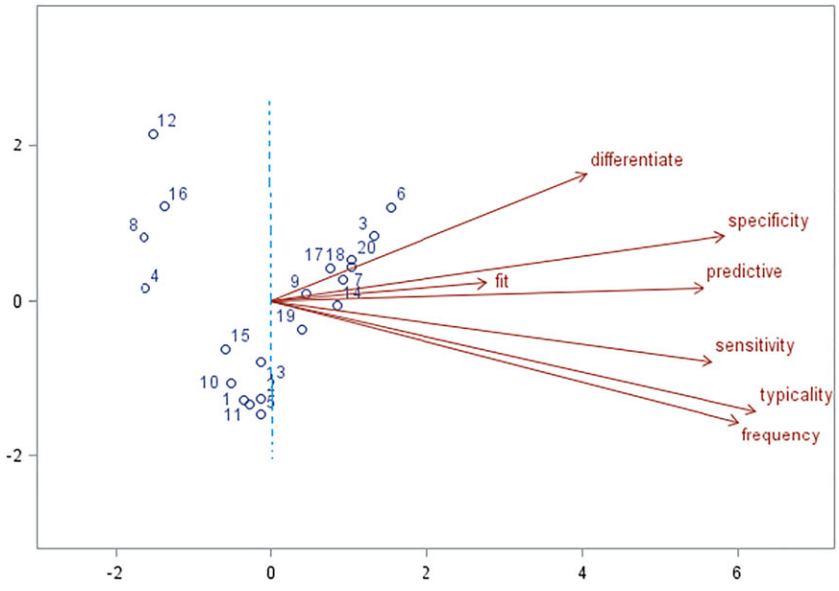

FIGURE 3 Item property fit. Relative length of the property vector indicates fit of item property to the configuration of items. One item property (i.e., fit) did not fit the configuration well in comparison with other item properties. This result is consistent with lower interrater reliability for item property of fit

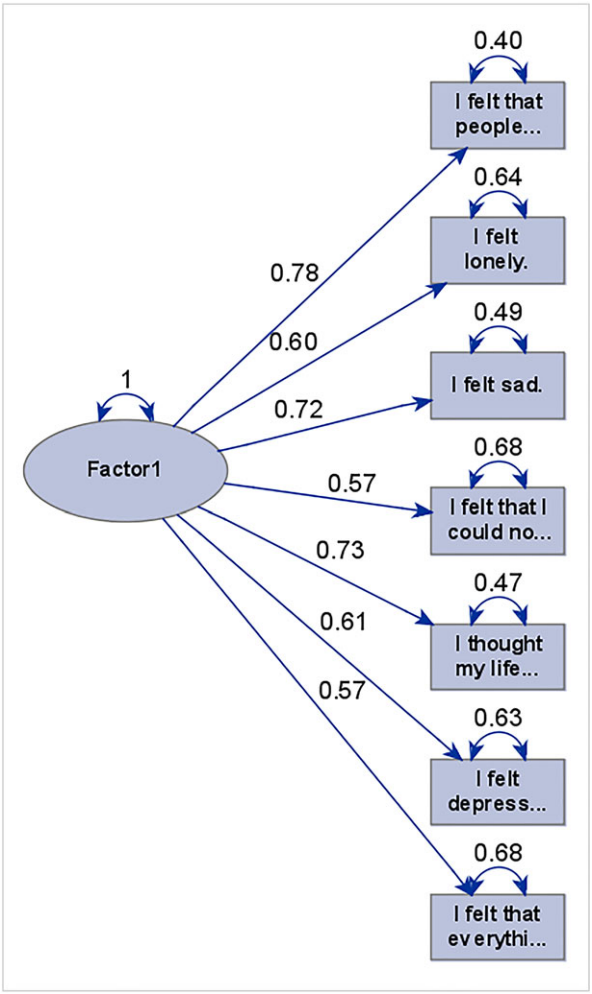

FIGURE 4 Factor structure of nine-item Center for Epidemiologic Studies Depression scale

6 | DISCUSSION

The purpose of this study was to propose a quantitative method for evaluating item content validity via item property analysis. Content validity is the first step in constructing a reliable and valid instrument. This step is crucial because the subsequent analyses are all hinged on this initial item pool. An initial item pool that is appropriately constructed can ensure improved reliability and validity. As it stands now, there is no quantitative method to analyze item content validity based on explicit criteria. In this study, we suggest a process of quantitatively examining the item content validity rather than using a holistic approach.

Methodologically, in the current practice of evaluating content validity, a group of experts are typically asked to judge the appropriateness of items without providing them with any explicit criteria as to what would constitute a good set of items. Thus, we are often left in the dark as to how the experts appraise the items and what guidelines they use in such an effort.

Given this black-box phenomenon of evaluating item content validity, we advocate that explicit criteria of item property should be used. To assess the property of items in the instrument, criteria for what constitutes good item properties should be clearly articulated when developing the item pool. In this study, we demonstrated how to construct a matrix of item properties that can be used by experts or individuals when judging the item quality. We also performed property fitting analysis to quantify the results. Specifically, we found that 12 counselors provided similar ratings across all items using the matrix of item properties, indicating that such a matrix may actually be viable. Moreover, these counselors perceived or interpreted the item properties in the same way, and no idiosyncratic biases were introduced in the rating process. These findings provided empirical evidence for supporting the use of explicit criteria of item property in appraising item content validity.

The result of property fitting analysis revealed that nine out of 20 CES-D items were determined to possess the seven item properties. Four positively worded items were located in the opposite direction, forming their own cluster. This pattern of item configuration was consistent with previous findings on positively and negatively worded items (e.g., Borgers, Hox, & Sikkel, 2004; Marsh, 1986); that is, they tended to form a separate factor. Results of exploratory analysis indicated a clear one-factor structure with high item factor loadings. The reliability estimate of scores from a nine-item version of the CES-D scale was higher than the full version of the scale based on this sample. These findings may suggest that using explicit criteria of item property may improve the quality of the instrument, at least in terms of reliability and factor structure.

6.1 | Implication

The proposed approach of evaluating item content validity may have some practical implications on instrument development. First, the method suggested here provides a way for clinicians and researchers to develop their measuring instruments in more planned fashion. That is, it forces the instrument developers to think more clearly about the nature of the instrument (e.g., diagnostic or predictive), and the instrument can be in the direction that maximizes the validity. We cannot simply construct "good item content validity" in isolation. We must design an assessment from the very start around the inferences we want to make, the situations that will evoke these inferences, and the series of reasoning that connects them.

Second, the approach discussed here can be used in any field of research where studies involve any survey or measurement instruments. Even in opinion survey studies, we can still benefit from items that are based on the explicit item properties that match our goals.

6.2 | Limitations

There are some limitations that must be noted. First, there is no universal set of explicit criteria of item property in evaluating items. Specific sets of explicit criteria of item property need to be developed based on the nature of the instrument. Second, we suggest the three kinds of quantitative methods in the process of evaluating item with respect to content validity. IRR and rater bias analysis are necessary so that we have confidence that the explicit criteria of item property are used by experts in a homogeneous way. Then we suggest to use MDS for property fitting analysis because it naturally fits into such an investigation. There may be other quantitative methods that can be equally effective. In this paper, we only suggest one of several potential methods.

Despite these limitations, this is the first study to suggest use of quantitative methods to evaluate content validity via explicit criteria of item properties. Although more efforts are needed in using this approach to develop a set of initial items on the front end, we believe that it can ultimately improve the item quality and save more time and resources on the back end. No matter who the experts or individuals will be when judging the appropriateness of items with regard to content validity, researchers should have a clear idea of how the items are appraised and what is used in such an evaluation.

DECLARATION OF INTEREST STATEMENT

The author has no competing interests.

ORCID

Cody Ding https://orcid.org/0000-0002-2894-1545

REFERENCES

Beauchamp, M. R., Barling, J., Li, Z., Morton, K. L., Keith, S. E., & Zumbo, B. D. (2010). Development and psychometric properties of the transformational teaching questionnaire. *Journal of Health Psychology*, *15*, 1123–1134. https://doi.org/10.1177/1359105310364175

Beck, C. T. (1999). Content validity exercises for nursing students. *Journal of Nursing Education*, *38*(3), 133–135.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point. *Quality & Quantity*, *38*, 17–33. https://doi.org/10.1023/B:QUQU.0000013236.29205.a6

Burns, J. R., & Rapee, R. M. (2016). Screening for mental health risk in high schools: The development of the Youth RADAR. *Psychological Assessment*, *28*, 1220–1231. https://doi.org/10.1037/pas0000237

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*, 238–319.

Chao, R. C.-L., & Green, K. E. (2011). Multiculturally Sensitive Mental Health Scale (MSMHS): Development, factor analysis, reliability, and validity. *Psychological Assessment*, *23*, 876–887. https://doi.org/10.1037/a0023710

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Data Theory Scaling System Group. (n.d.). *PROXSCAL* (Version 1.0). Leiden university, Netherlands: Faculty of Social and Behavioral Sciences.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Ding, C. (2018). Fundamentals of applied multidimensional scaling for educational and psychological research: Springer.

Draheim, C., & Harrison, T. L., Embretson, S. E., & Engle, R. W. (2017, March 9). What item response theory can tell us about the complex span tasks. psychological assessment, Advance online publication. https://doi.org/10.1037/pas0000444

Faulstich, M. E., Carey, M. P., Ruggiero, L., Enyart, P., & Gresham, F. (1986). Assessment of depression in childhood and adolescence: An evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). *American Journal of Psychiatry*, *143*(8), 1024–1027. https://doi.org/10.1176/ajp.143.8.1024

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*(1), 22–36.

Funk, B. A. III, Huebner, E. S., & Valois, R. F. (2006). Reliability and validity of a brief life satisfaction scale with a high school sample. *Journal of Happiness Studies*, *7*, 41–54.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorial in Quantitative Methods for Psychology*, *8*(1), 23–34.

Horan, C. B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structure. *Psychometrika*, *34*, 139–165.

Hughes, L., Betka, S., & Longarzo, M. (2018). Validation of an electronic version of the Self-Awareness Questionnaire in English and Italian healthy samples. *International Journal of Methods in Psychiatric Research*, First published: 03 December 2018. https://doi.org/10.1002/mpr.1758

IBM Corp (2017). *IBM SPSS Statistics for Windows (Version 25)*. Armonk, NY: IBM Corp.

Kanter, J. W., Mulick, P. S., Busch, A. M., Berlin, K. S., & Martell, C. R. (2007). The Behavioral Activation for Depression Scale (BADS): Psychometric properties and factor structure. *Journal of Psychopathology and Behavioral Assessment*, *29*(3), 191–202.

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H. U. (2006). The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). International Journal of Methods in Psychiatric Research, First Published: 24 March 2006, 171–185.

Klassen, R. M., Usher, M. B., Ellen, L., Chong, W. H., Huan, V. S., Wong, I. Y. F., & Georgiou, T. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries. *Contemporary Educational Psychology*, *34*(1), 67–76.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27.

Lehmann-Willenbrock, N., & Kauffeld, S. (2010). Development and construct validation of the German Workplace Trust Survey. *European Journal of Psychological Assessment*, *26*(1), 3–10.

Lenz, A. S., Balkin, R. S., Gómez Soler, I., & Martínez, P. (2015). Development and evaluation of a Spanish-language version of the relational health indices. psychological assessment, ADVANCE online publication. https://doi.org/10.1037/pas0000170

Lu, L., & Gilmour, R. (2006). Individual-oriented and socially oriented cultural conceptions of subjective well-being: Conceptual analysis and scale development. *Asian Journal of Social Psychology*, *9*, 36–49. https://doi.org/10.1111/j.1367-2223.2006.00183.x

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*, 539–549.

Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2015). What range of trait levels can the autism-spectrum quotient (AQ) measure reliably? An item response theory analysis. Psychological assessment, advance online publication. https://doi.org/10.1037/pas0000215

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 380–401.

Rosenberg, M. (1965). *Society and adolescent self-image*. Princeton, NJ: Princeton University Press.

SAS. (2013). Base SAS® 9.4 procedures guide: Statistical procedures (2 ed.). Cary, NC: SAS Institute, Inc.

Schellingerhout, J. M., Verhagen, A. P., Heymans, M. W., Koes, B. W., & de Vet, H. (2012). Measurement properties of disease-specific questionaires in patients with neck pain: A systematic review. *Quality of Life Research*, *21*, 659–670.

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology*, *5*(509). https://doi.org/10.3389/fpsyg.2014.00509

Storch, E. A., Rasmussen, S. A., Price, L. H., Larson, M. J., & Murphy, T. K. (2010). Development and psychometric evaluation of the Yale–Brown Obsessive-Compulsive Scale—Second Edition. *Psychological Assessment*, *22*(2), 223–232.

Terwee, C. B., Bot, S. D. M., Boer, M., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionaires. *Journal of Clinical Epidemiology*, *60*, 34–42.

Walker, D. A. (2010). A confirmatory factor analysis of the attitude toward research scale. *Multiple Linear Regression Viewpoints*, *36*(1), 17–24.

Williams, S. L., & Polaha, J. (2014). Rural parents' perceived stigma of seeking mental health services for their children: Development and evaluation of a new instrument. *Psychological Assessment*, *26*, 763–773.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4*(2), 165–4178. https://doi.org/10.15171/jcs.2015.017

How to cite this article: Ding C. Examining item content validity using property fitting analysis via multidimensional scaling. *Int J Methods Psychiatr Res*. 2019;28:e1771. https://doi.org/10.1002/mpr.1771

APPENDIX A

MATRIX OF ITEM PROPERTY FOR CES-D SCALE

CES depression scale

Item Property Criteria:

Typicality—how typical this symptom is for depression?

Frequency—how often this symptom occurs in depression?

Differentiating—can this symptom differentiate individuals with depression from those without?

Sensitivity—is this symptom sensitive to depression?

Specificity—is this symptom specific to depression?

Fit—is this symptom appropriate to the general population?

Predictive— can this symptom predict the depression?

Rating:

Based on above description of these criteria, please rate each item on a scale of **0 to 9**, with **0** indicating "*not a good item*" with respect to these criteria and **9** indicating "*a good item*".

	Typicality	Frequency	Differentiating	Sensitivity	Specificity	Fit	Predictive
I was bothered by things that usually don't bother me.							
I did not feel like eating; my appetite was poor							
I felt that I could not shake off the blues even with help from my family or friends.							
I felt I was just as good as other people.							
I had trouble keeping my mind on what I was doing.							
I felt depressed.							
I felt that everything I did was an effort.							
I felt hopeful about the future.							
I thought my life had been a failure.							
I felt fearful.							
My sleep was restless.							
I was happy.							
I talked less than usual.							
I felt lonely.							
People were unfriendly.							
I enjoyed life.							
I had crying spells.							
I felt sad.							
I felt that people dislike me.							
I could not get "going."							