2020 ■ Vol. 16 ■ no. 4

Visualizing Items and Measures: An Overview and Demonstration of the Kernel Smoothing Item Response Theory Technique

Gordana Rajlic [a]

[a]University of British Columbia

Abstract ■ The current demonstration was conducted to familiarize a broader audience of applied researchers in psychology and social sciences with the benefits of an exploratory psychometric technique – kernel smoothing item response theory (KSIRT). A data-driven, nonparametric KSIRT provides a visual representation of the characteristics of the items in a measure (scale or test) and offers convenient preliminary feedback about the functioning of the items and the measure in a particular research context. The technique could be a useful addition to the analytical toolkit of applied researchers that work with a range of measures, within the classical test theory or IRT framework. KSIRT is described and its use is demonstrated with a set of items from a psychological well-being measure. A recently developed, easy to use R package was utilized to perform the analyses and the R code is included in the manuscript.

Keywords ■ exploratory psychometric analysis; IRT; kernel smoothing; nonparametric regression; visualization. **Tools** ■ R.

grajlic@gmail.com

10.20982/tqmp.16.4.p363

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers ■ Sébastien Béland (Université de Montréal)

Introduction

The purpose of the current project was to convey more information to applied researchers about kernel smoothing item response theory (KSIRT; Ramsay, 1991, 2000) — an exploratory, data-driven technique that provides graphical displays of the item-level functioning of a measure. KSIRT graphs facilitate an understanding of the measures that researchers use in their work. They can point to problems with the functioning of the items and the measure in a particular research context (e.g., violations of the measurement assumptions such as monotonicity assumption or item invariance assumption) that may impact the research findings and conclusions. KSIRT was described in the current project, and its application was demonstrated with a set of items from a psychological well-being measure.

KSIRT belongs to a class of non-parametric item response theory (IRT) techniques. In the IRT framework, item responding is modeled as an interaction between the latent dimension, representing the construct the measure was intended to measure, and the characteristics of the items (see Baker, 2001; de Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; van der Linden & Hambleton, 1997). As a model-based measurement, IRT has offered improved solutions for various practical measurement problems in psychology and social sciences (Embretson & Reise, 2000). A great number of IRT models have been developed, and among those models, a distinction could be made between parametric IRT and nonparametric IRT models. In both parametric and nonparametric approaches, the relation between the latent dimension and the probability of a "correct" (or positively keyed) item response is described by a monotonically increasing function; this function – item response function, is graphically represented by an item characteristic curve (ICC). In parametric IRT models, a specific item response function is used (such as logistic function) and ICCs have a prespecified form, with the models differing in how the ICC is characterized – what function is used and how many item parameters are proposed to determine the ICC (Birnbaum, 1968; Bock, 1972; Lord, 1952;

Lord & Novick, 1968; Rasch, 1960; Samejima, 1969). In nonparametric IRT models, ICCs are estimated without imposing a specific parametric function, that is, without assuming the particular form of the curve (Douglas, 1997; Junker & Sijtsma, 2001; Molenaar, 1997; Mokken, 1971; Molenaar, 1997; Ramsay, 1991; Samejima, 1998; Sijtsma & Molenaar, 2002). Overall, parametric IRT models are based on "strong" assumptions, whereas in nonparametric models some of the strong assumptions of parametric models are relaxed, such as the assumption about the particular shape of the ICC curve. Among most used nonparameteric IRT models are the Mokken model (Mokken, 1971; Molenaar, 1997) and KSIRT (Ramsay, 1991). In KSIRT, ICCs are estimated from data by using kernel smoothing nonparametric regression (Eubank, 1988; Härdle, 1990) and the resulting ICCs are a direct reflection of the data at hand. Further technical details about a parametric IRT model and nonparametric KSIRT procedure are provided in Appendix.

KSIRT provides a different type of information compared to the information resulting from fitting a parametric IRT model. The ICCs from KSIRT are data-driven visual representations of the relations between the latent dimension and item responses, and they provide a convenient preliminary feedback to the researchers about the characteristics of individual items and the measure in the given application (e.g., they provide information about item discrimination across different levels of the latent trait, about monotonicity assumption, and when plotted for different groups of respondents, they could indicate a need for further group analyses such as measurement invariance analyses, as described in the following sections). Some of the practical uses of KSIRT and other nonparametric IRT techniques were presented in Meijer and Baneke (2004), Santor, Ramsay, and Zuroff (1994), and Sijtsma, Emons, Bouwmeester, Nyklíček, and Roorda (2008), whereas methodological research concerning the use of KSIRT include Douglas (1997), Douglas and Cohen (2001), Wells and Bolt (2008).

KSIRT could be a useful addition to both the traditional classical test theory (CTT) item analysis and to the parametric IRT analyses. In relation to CTT, in which the main focus is on the total score and the measure as a whole, KSIRT contributes by bringing focus on the individual items (i.e., item-level functioning of the measure). It assists in assessing the item characteristics and identifying possible item-level measurement issues. In relation to the CTT-based item analysis (e.g., item-total correlations and item difficulty indexes), as opposed to numerical summaries, KSIRT brings data-driven visual presentations of item characteristics, as well as the information that is not available in CTT analysis. For example, information about the changes in item discrimination across different levels of the trait is not obtained in the CTT item analysis. In relation to the parametric IRT models, some of the proposed benefits of nonparametric KSIRT include its use in the evaluation of the monotonicity assumption of the parametric models, in guiding the selection of the appropriate parametric model (if a one-, two-, or three parameter IRT parametric model may be appropriate to the data), or in assessment of the model fit (Douglas, 1997; Douglas & Cohen, 2001; Lee, Wollack, & Douglas, 2009; Sueiro & Abad, 2011; Wells & Bolt, 2008). Due to a different estimation methods, nonparametric IRT can be used with a smaller number of respondents and items compared with what is needed for parametric IRT[1], as emphasized in Junker and Sijtsma (2001) and Stout (2001).

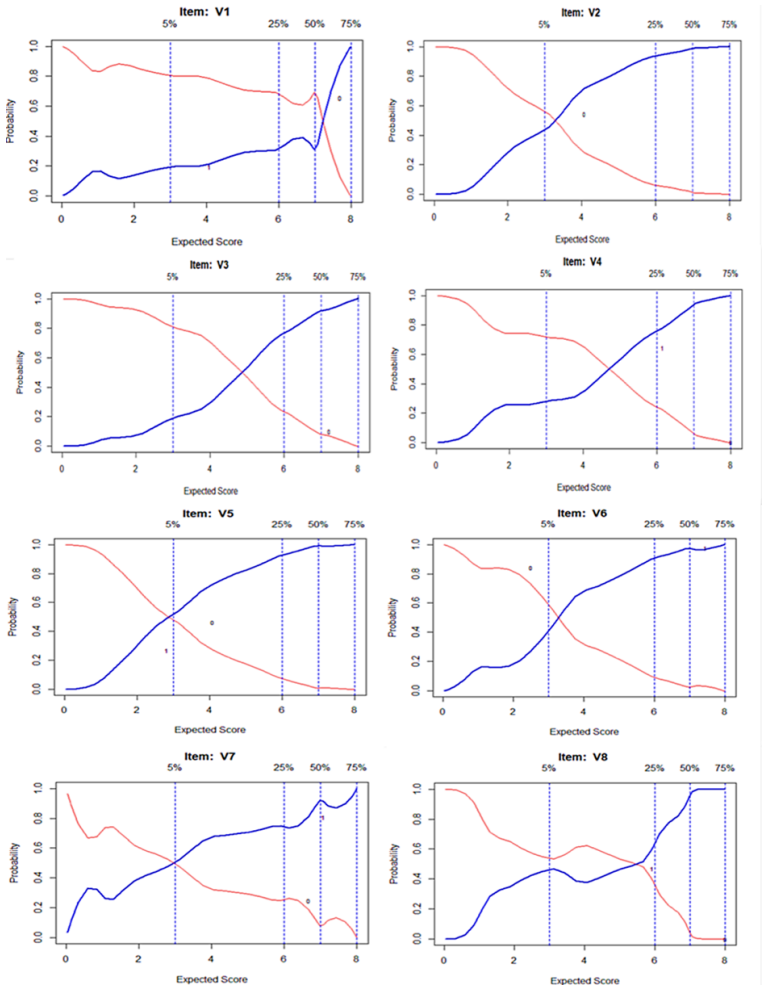
KSIRT has not been used often in applied research practice – one of the reasons for its infrequent use is researchers' non-familiarity with the technique, with its practical application and interpretation of the results. To convey more information about this technique to a broader range of researchers that work with psycho-social measures was a motivation for the current project. Other reason for the infrequent use of the technique may have been a need for the use of the specialized software (KSIRT has been traditionally conducted by using Ramsay's TestGraf software; Ramsay, 2000). A convenient R package (R Core Team, 2019) has been developed recently – KernSmoothIRT (Mazza, Punzo, & McGuire, 2014). As the use of R has been becoming regular among social science researchers, the work of Mazza et al. (2014) may initiate a greater use of the technique.

An Application of KSIRT

As a demonstration, KSIRT technique was used with a set of items from a psychological well-being measure (the Scales of Psychological Well-Being; Ryff, 1995; Ryff & Keyes, 1995). Specifically, the items that assess the "purpose of life" dimension of well-being and comprise the Purpose in Life Scale (PL Scale) were used. The focus of the demonstration was on the description of the information that researchers could obtain from a KSIRT application – the elements of the KSIRT graphs and the information conveyed by the graphs were described in detail. The procedure and the results are presented in non-technical terms, with a further description of the main steps of the procedure included in Appendix. A detailed technical description of KSIRT and

[1] In parametric IRT approaches, larger sample sizes are required for the estimation; with increase in the number of parameters in different models, larger samples are needed. For a two-parameter model, for example, a minimum required sample size of about 500 respondents was recommended (DeMars, 2010). KSIRT can be used with smaller samples; for instance, a sample of about 200 was used in Ramsay (2000).

Figure 1 ■ Option characteristic curve (OCC) graphs for the eight PL Scale items - with the expected total score on the x-axis. The blue curve is for the positively keyed item response (scored 1), and the red curve is for the opposite response (scored 0).



nonparametric regression technique can be found in Eubank (1988), Härdle (1990), Mazza et al. (2014), and Ramsay (1991).

The eight PL Scale items are statements about the sense of purpose of life, with higher scores on the PL Scale indicating that the respondent "has goals in life and a sense of directedness, feels there is meaning to present and past life, holds beliefs that give life purpose, has aims and objectives for living" (Ryff, 1995).[2] The PL Scale data used in this demonstration were from the United States National Health Measurement Study, 2005-2006 (Fryback, 2009), from an open access database. The data from 3680 participants (1578 males and 2102 females) who provided answers to all PL Scale items were used. Dichotomously scored items, with a score of 1 representing agreement with the statement and the score 0 representing disagreement with the statement, were analyzed, with a total PL Scale score ranging from 0 to 8. The negatively worded items were recoded. The mean total score was 6.4 ($SD = 1.6$), with the item endorsement proportions ranging from .51 through .91. In the given sample, the ordinal coefficient alpha was .80. KSIRT was applied to the PL Scale and the item curves for the eight items were plotted. In addition to the KSIRT analysis in the whole sample, the analyses were

[2]For further detail about the scale and the specific content of the items, refer to Ryff (1995) and Ryff and Keyes (1995). The items and the scale have been scored in different ways in past studies; in the current demonstration, dichotomous item scoring was used.

performed for the specific subsamples (gender groups in this case) and the group item-responding patterns were presented. To make it easier for researchers to perform the analyses described in this project, a link to the subset of the variables used for the demonstration was provided (included in Appendix).

The Elements of the Option Characteristic Curve (OCC) Graphs

An application of KSIRT to the PL Scale, in the particular sample of respondents, resulted in the option characteristic curve (OCC) graphs presented in Figure 1. Because the curves for the two response options (coded 0 and 1) were presented in the graphs, instead of the term "ICC", a more correct term "OCC" is used.[3] On the OCCs graphs, the y-axis is the probability of a certain response, ranging from 0 to 1, whereas the x-axis represents the latent dimension on which the respondents are ordered (i.e., the construct the measure was designed to measure). For each of the PL Scale items, two curves were plotted. Displayed in blue (darker) color in the graphs is the curve representing the relation between the latent dimension and the probability of the correct response (the response coded 1, indicating agreement with the purpose in life statement). The other curve visualizes the response coded 0; accordingly, this curve decreases from left to right with increase in the latent dimension. The curve for the response 0 is just the opposite of the curve for the response 1 – in the case of polytomous items other response options of interest to the researchers can be visualized and multiple curves plotted. A display variable on the x-axis is expected scores (the number of items that a respondent at a particular position on the latent dimension would, on average, endorse), ranging from 0 to 8. Latent dimension can be scaled in different ways (Mazza et al., 2014; Ramsay, 2000). Different types of scaling are more convenient in different situations: When the focus of the analysis is on a specific measure and its properties, plotting the expected scale scores on the x-axis may be more informative to researchers, whereas when the trait in general is of interest (i.e., higher level of generalization), the standard scores scaling is more convenient.

In the presented OCC graphs, there are dashed vertical lines – these five lines indicate the points below which 5%, 25%, 50%, 75%, and 95% of the respondents fall in terms of actual total score. The positons of the vertical lines are determined on the measure-level, that is, the position of the lines is the same for all the items in the measure. In the OCC graphs in Figure 1, the 25% line was placed at the score 6, indicating that 25% of the respondents fall below the score 6. In other words, 75% percent of respondents were in the range of scores from 6 to 8. This demonstrates negative skewness of the observed scale scores, indicating an "easy" measure/test, in which the majority of the respondents obtained high scores (agreed with the purpose of life statements).

Inspecting the Shape of the Curve

The OCC's convey information about certain characteristics of items and there are several aspects of the curve that should be examined. First, it is useful to examine whether there is an increase in probability to endorse an item with increase along the latent dimension, which is an assumption of the most often used measurement models (the monotonicity assumption). Traditional monotonic measurement models assume that item endorsement is increasingly likely with an increase in the latent trait/ability. In IRT, item response function specifies that examinees with higher scores on the trait have higher expected probabilities for answering the item correctly than examinees with lower scores on the traits (Hambleton & Swaminathan, 1985). An ICC that satisfies monotonicity assumption should increase from left to right;[4] therefore, violations of monotonicity are manifested as a curve decreasing from left to right, a curve in a form of a "wave", a "U-shaped' curve", etc. Assumption of monotonicity must be investigated because violations of monotonicity affect the accuracy of measurement when monotonic psychometric models are used (Hambleton & Swaminathan, 1985; Sijtsma & Molenaar, 2002).

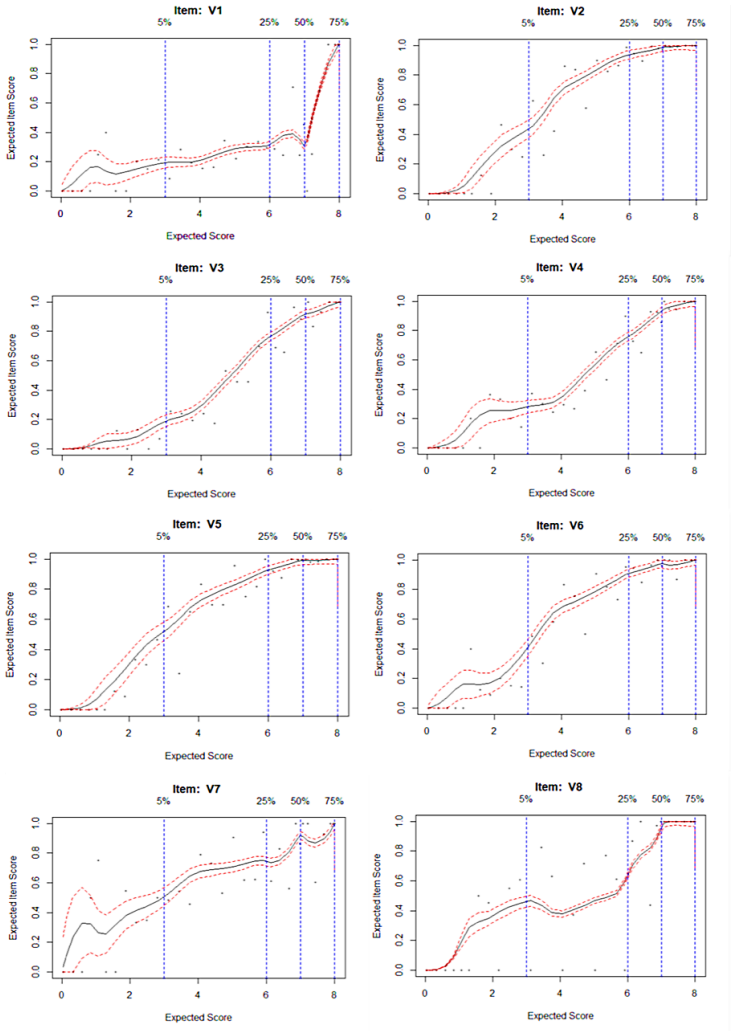
In reviewing monotonicity, it is also useful to plot 95% confidence intervals for the value of the curve at specific latent dimension values (pointwise confidence intervals). The expected item score graphs for the PL Scale items with 95% confidence intervals are presented in Figure 2. The intervals tell us how precisely the curve has been estimated at specific levels of the latent dimension, which is dependent on the number of respondents. For the PL Scale items, the widest regions are at the low end of the latent dimension (low purpose in life), where the smallest number of respondents fall (e.g., on the left from the 5th percentile line). There is less data for estimating the curve in this region and, consequently, there is less precision in the estimates.

An inspection of the curves plotted for the PL Scale items suggests a certain degree of violation of monotonicity in item 8, as well as minor distortions of monotonicity in items 1 and 7 (violation in the form of a shallow, narrow-range dip). Items with distortions of monotonicity should

[3] OCC are item response curves plotted for more than one item response option, for example, five curves could be plotted for items with five response options. The term ICC is used for a single curve, plotted for a correct/positively keyed response (i.e., for a single response option).

[4] That is, it should not decrease: The probability of response is "non-decreasing" function of the latent trait in nonparameteric IRT.

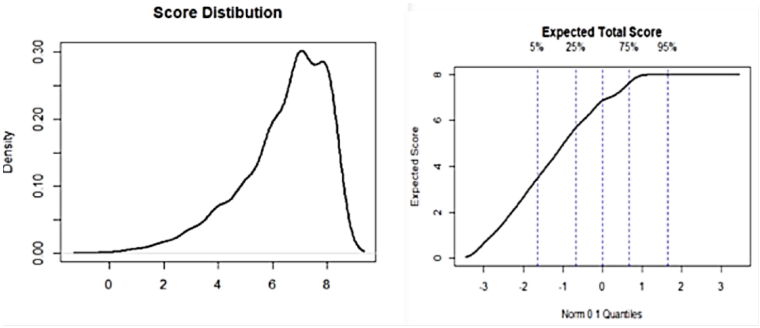
Figure 2 ■ Expected item score (EIS) graphs for the eight PL Scale items with corresponding 95% pointwise confidence intervals (with the expected total score on the x-axis and the expected item score on the y-axis).



be flagged for further examination. Other characteristics of the items and the measure should be looked at to understand what the problem may be and what course of action to take. Minor violations of monotonicity may be tolerated if the item has clear benefits for the scale in some other respect (Sijtsma & Meijer, 2006). Serious distortions in monotonicity (e.g., a U-shaped curve or an inverse U-shaped curve), however, are problematic, and items with such distortions indicate problems with the functioning of the measure. They could suggest that different measurement models, for non-monotone item response functions, may be more appropriate (Sijtsma & Meijer, 2006), or in a particular situation, they can point to the violations of the other assumptions and to the measurement invariance issues.

Two characteristics of the items that are commonly examined in the item analysis (CTT or IRT based), item location and item discrimination, can be conveniently inspected from the plotted OCC graphs. Item location, also known as item threshold or item difficulty, is the location on the latent dimension at which the probability of a correct response (or endorsing the item) is 0.50. That is, at the trait levels higher than the threshold, the probability of endorsing the item becomes higher than the probability of not endorsing the item. In the graphs in Figure 1, item location is the value on the x-axis corresponding to the point at which the two plotted curves cross. Item location increases from left to right: Items located at the low end of

Figure 3 ■ (a) Kernel density estimate of the distribution of the total score (b) Expected total score as a function of the quantiles of the standard normal distribution.



the latent dimension are characterized as easy items (they are endorsed by respondents at the lower end of the latent dimension) and items with a high item location are characterized as difficult items. Based on the OCC plots, the items can be easily compared visually and ordered in terms of their item locations. For example, based on reviewing the graphs in Figure 1, the easiest PL Scale items were items 5 and 7, whereas the item with the highest item location was item 1. The PL Scale items, overall, can be characterized as easy in the given sample of respondents, with most of the items located at the lower end of the latent dimension, consistent with the negative skewness of the scale scores.

Another important aspect of the ICC curve is its steepness. The steepness of the curve represents item discrimination and it indicates how rapidly the probability of correct response changes with changes along the latent dimension. The steeper line indicates that the item is good at distinguishing between respondents with lower and higher levels of trait; that is, the steeper line indicates higher discrimination whereas the flatter line indicates lower discrimination and possibly problematic items. Based on the KSIRT graphs, items can be visually compared in their discriminative power. Additionally, because the shape of the curve in KSIRT is not prespecified (as in parametric IRT models), KSIRT enables researchers to review the changes in item discrimination along the latent dimension, and to visually compare items in terms of their discrimination at different regions of the latent dimension. In the relevant parametric IRT (e.g., two- or three-parameter models), a single item discrimination index is provided, with this value proportional to the slope of the ICC curve at the point of item threshold/difficulty).[5]

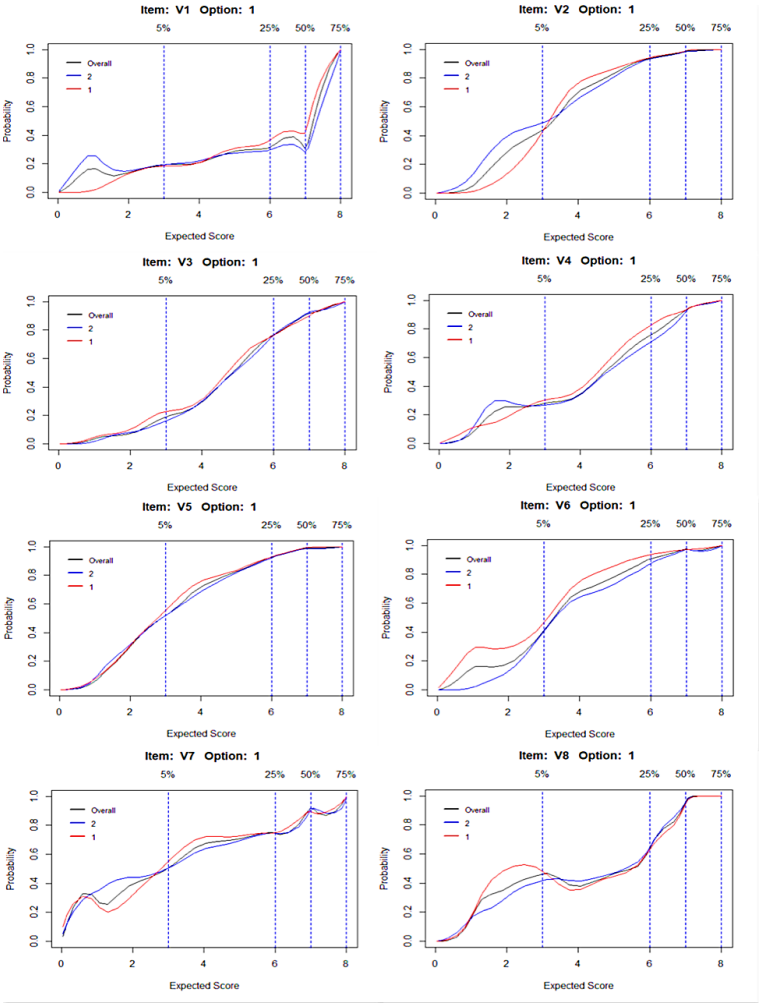
The differences among the items in terms of item discrimination can be observed in the OCC graphs for the PL Scale items. For example, item 1 demonstrates poor discrimination across the lower/middle regions of the latent dimension, but is highly discriminative at the higher end of the dimension, as opposed to item 2 that is most discriminative at the lower end and is not discriminative at the higher end of the dimension (indicated by the flat curve at this end of the dimension). The graphs suggest that the majority of the PL Scale items discriminate poorly at the higher end of the latent dimension. In this context, item 1 may be beneficial, despite its poor discrimination at the lower/middle regions of the latent dimension. The intended use of the measure and the target population must be kept in mind in interpreting the KSIRT results.

The investigation of the shape of the curves should also include an examination of the lower and upper asymptote of the curve (i.e., the lowest and the highest end of the curve). The probability of endorsing an item should approach 1 in the highest region of the latent dimension and should approach 0 in the lowest region of the latent dimension. This was the case with all eight PL Scale items (Figure 1). If "non-one" higher asymptote and "non-zero" lower asymptote were observed, this information would suggest that issues such as guessing or social desirability may be of relevance in item responding and that they should be taken into account. Such information is useful in parametric modeling and it assists in choosing the appropriate parametric model (e.g., in deciding if three- or four-parameter models would be more appropriate than a two-parameter model).

In addition to the item-level graphs, some measure-level summary plots can be plotted. The probability density graph (kernel density estimate of the distribution of the total score), indicating how probable scores are by the height of the function, is presented in Figure 3a, showing

[5]Item discrimination index in the two-parameter IRT model is related to factor loadings in item FA because of the mathematical relation between the two-parameter IRT models and the FA of item responses (Wirth & Edwards, 2007).

Figure 4 ■ Option characteristic curve (OCC) graphs for the eight PL Scale items for males (the curve in red, lighter color) and females (the curve in blue, darker color).

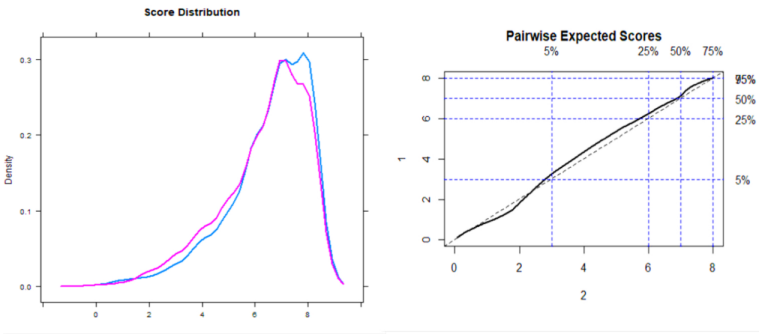


that the PL Scale scores in the 6 to 8 range are most probable. The test characteristic curve graph that depicts the expected PL Scale score in relation to the standard normal distribution (as a function of the quantiles of the standard normal distribution) is presented in Figure 3b, showing if the monotonicity requirement is met at the test level. The graphs visualize skewness in the PL Scale scores.

OCCs in Two Groups of Respondents

If certain groups of responders are of particular interest, the analyses can be conducted to assess if the items function in a similar way across the relevant groups. The differences in the shape of the curves and the size of the areas between the curves for different groups could point to the issue of differential item functioning (DIF) and measurement invariance in the measure (Holland & Wainer, 1993; Lord, 1980; Meredith, 1993; Raju, 1988). DIF is present when respondents from different groups have differing probabilities to endorse the item at the same level of the trait/ability of interest. Measurement invariance is a more general concept that can refer to both item level and measure level invariance. The difference is usually made between uniform and nonuniform DIF, with uniform DIF referring to the type of DIF when the probability to endorse an item is higher in one group across levels of the trait/ability, whereas in non-uniform DIF the probability to endorse an item is higher for one group at certain levels of the trait and higher for the other group at other levels of the trait. That is, uniform DIF occurs when there is no substantial interaction between the trait level and group mem-

Figure 5 ■ (a) Total score distribution plot – kernel density estimate of the distribution of the observed scores for females (red line) and for males (blue line); (b) The pairwise expected score plot for the two groups, with females on the x-axis and males on the y-axis.



bership, whereas nonuniform DIF implies interaction between the trait level and group membership (Mellenbergh, 1982; Zumbo, 2007).

To demonstrate the DIF analyses with the PL scale items, the OCCs were plotted in two different groups; specifically, the curves were plotted for male respondents and for female respondents, along with the overall curve for all respondents (Figure 4). In regard to item 3 and 5, for example, the curves are very similar and close to each other, suggesting that there are no differences in the functioning of these items in the two groups. The curves for item 2 indicate that, at the lower end of the purpose of life dimension, the probability to agree with the statement is higher for females compared to males, whereas at the middle end of the latent dimension, the probability to agree was somewhat higher for males, suggesting possible nonuniform DIF. In relation to item 6, the probability of agreeing with the statement was greater in males along most of the latent dimension (i.e., at different ends of the latent dimension), suggesting uniform DIF. Overall, visual inspection of the item curves obtained for the relevant groups provides preliminary information about items functioning in those groups and it can suggest a need for more detailed analyses. When the differences are noted, further analyses are recommended.

In addition to the item curves, some measure-level graphs are also available for different groups. The distributions of the scores in the groups of males and females are presented in Figure 5a whereas the distributions of the scores in the two groups are compared in a Q-Q plot in Figure 5b. Greater deviation from the diagonal line in Figure 5b[6] indicates greater differences in the two distributions; in this case, the deviation from the diagonal line was slight. The graphs do not suggest substantial differences in the distributions of the scores between the two groups.

In relation to the DIF analyses, many methods and techniques have been proposed for assessing DIF, such as the Mantel-Haenszel procedure, logistic regression, SIBTEST, and IRT based procedures (see Millsap, 2011). Research is ongoing about the utility of these methods in different situations and about the conditions of their use. For a difference from the parametric IRT methods that provide a summary index for DIF (e.g., statistical indexes about when the difference between the curves is statistically significant, and various effect size indices), KSIRT does not provide a single summary index for DIF. Rather, it provides data-driven (as opposed to model-driven) ICC curves that offer a visual presentation of item responses in different groups. They could point to DIF, both uniform and nonuniform, and reviewing the KSIRT curves is a good starting point for further, in-depth examination of DIF issues, with various other methods of choice.

Discussion

KSIRT graphs provide visual information about item-level measure functioning in the given measurement context. Different from parametric IRT models, in which the relationship between item responses and a latent trait is represented by a specific mathematical function (e.g., logistic or normal ogive), in nonparametric KSIRT the form of the function is not specified; therefore KSIRT graphs are a more direct reflection of the data at hand. The graphs provide convenient preliminary feedback about whether monotonicity assumption is met, item discrimination along the different regions of the latent dimension, items functioning in different groups of respondents, and possible

[6] A dotted diagonal line is plotted as a reference line. When performance of the two groups does not differ, the relation is plotted as an approximately diagonal line.

measurement invariance issues. The KSIRT analyses supplement traditional CTT and IRT parametric analyses, and they can be used for different purposes and in different research contexts (Douglas & Cohen, 2001; Junker & Sijtsma, 2001; Samejima, 1998; Sijtsma, 2005; Sijtsma & Molenaar, 2002; Stout, 2001). Ramsay (1991) emphasized the role of KSIRT as an exploratory data analysis tool – such a use of KSIRT was demonstrated in the current project.

Based on the real data from a well-being measure, the relevant KSIRT graphs were created, their elements described, as well as the information that the researchers could obtain from the graphs. KSIRT graphs of the PL Scale indicated overall well-behaved items and no obvious problems in the measure functioning in the particular research context. Gender differences were noted in the functioning of some of the PL Scale items, however, they did not seem to affect the overall performance of the scale in the two groups. When problematic item functioning is suggested by KSIRT graphs (e.g., monotonicity violation, low discrimination, possible measurement invariance problems), further psychometric analyses should be conducted, and the problems investigated. Such analyses and the possible course of action in regard to the problematic items are important topics that were out of scope of the current project and were not addressed here.

Different from the parametric IRT models, KSIRT does not provide a single index of item discrimination, or test and item information functions estimates. There is no statistical index for DIF/measurement invariance, in relation to item functioning in the relevant groups. These are sometimes noted as the limitations of this approach. However, KSIRT provides different type of information, beneficial to researchers in a different way, as described in this demonstration. When point estimates are needed, other psychometric analyses should be utilized. Checking of the assumptions of unidimensionality and local independence (assumptions of KSIRT and parametric IRT modeling) was not performed in this report. Unidimensionality posits that a single latent dimension underlies the responses on the given measure, whereas the related assumption of local independence assumes that different items responses are independent conditioning on the underlying trait(s). The assessment of the two assumptions should be regularly performed before other IRT analyses are conducted, by using the techniques described in Hambleton et al. (1991); Hattie (1985); Zhang and Stout (1999).

In terms of software, the KSIRT analyses can be performed in freely available R, by using a convenient R package KernSmoothIRT (Mazza et al., 2014), or by using the traditionally used software – TestGraf (Ramsay, 2000), which can be requested from the author. For the analyses performed in this report, KernSmoothIRT was used, with the R code provided in the Appendix. For further details about the procedure, see Mazza et al. and Ramsay. In this report, KSIRT was conducted with dichotomously scored items (such as items with yes-no, present-not present, agree-not agree responses) in order to provide a clear and easy-to-understand example of the procedure. The understanding of the procedure, however, generalizes to polytomous items (with several response categories), with some of the analyses and interpretations more complex in such a context. The description and examples of the use of KSIRT with different types of items can be found in Mazza et al. (2014); Ramsay (2000), Santor et al. (1994); and such analyses can be performed with both KernSmoothIRT and TestGraph.

In conclusion, KSIRT could be a useful addition to analytical tools of applied researchers who use various psycho-social measures (tests or scales; self-report or performance/task based), which follow the assumption that the item responses reflect an underlying latent dimension (i.e., the target construct). KSIRT visual presentation of the characteristics of the items conveys to the researchers, in a convenient and easy-to-understand way, preliminary information about the functioning of the items and the measure in the particular research context, and can point to possible measurement problems. If, in the given context, problematic items' functioning is noted, the research results based on the use of the measure cannot be trusted, and further analyses are needed. KSIRT is one of the procedures that can be helpful to applied researchers in ensuring the quality of their research findings and conclusions. Additionally, a regular inclusion of KSIRT graphs in research reports (as a "visual guide" to items/measure) would aid in communicating the characteristics of the measures and of the corresponding research to the research community. The intention of the current demonstration was to present the issues of interest to a wider range of audiences; for technical details of the procedure, the original work of Ramsay (1991, 2000) and Mazza et al. (2014) should be consulted.

References

Baker, F. (2001). Basics of item response theory. Retrieved from http://echo.edres.org:8080/irt/baker

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. doi:10.1007/BF02291411

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C. (2010). *Item response theory*. New York: Oxford.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28. doi:10.1007/BF02294778

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234–243. doi:10.1177/01466210122032046

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker.

Fryback, D. G. (2009). *United states national health measurement study, 2005-2006* (tech. rep. No. ICPSR23263-v1). doi:10.3886/ICPSR23263.v1

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Härdle, W. (1990). *Applied nonparametric regression (econometric society monographs)*. doi:10.1017/CCOL0521382483

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–164. doi:10.1177/014662168500900204

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*, 211–220. doi:10.1177/01466210122032028

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*, 181–197. doi:10.1177/0013164408322026

Lord, F. M. (1952). *A theory of test scores (psychometric monograph no. 7)*. Richmond, VA: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mazza, A., Punzo, A., & McGuire, B. (2014). Kernsmoothirt: An r package for kernel smoothing in item response theory. *Journal of Statistical Software*, *58*, 1–34. Retrieved from http://www.jstatsoft.org/v58/i06/

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*, 354–368. doi:10.1037/1082-989X.9.3.354

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105–118. doi:10.2307/1164960

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In J. v. d. Linden & R. K. Hambleton (Eds.), *W* (pp. 367–380). Handbook of modern item response theory . New York: Springer.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from. Retrieved from https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. doi:10.1007/BF02294403

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630. doi:10.1007/BF02294494

Ramsay, J. O. (2000). Testgraf: A program for the graphical analysis of multiple-choice tests and questionnaire data. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Revelle, W. (2017). Using R for psychological research: A simple guide to an elegant language. Retrieved from http://personality-project.org/r/r.guide.html

Ryff, C. D. (1995). Psychological well-being in adult life. *Current Directions in Psychological Science*, *4*, 99–104. doi:10.1111/1467-8721.ep10772395

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69*, 719–727. doi:10.1037/0022-3514.69.4.719

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (psychometric monograph no. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika*, *63*, 111–130. doi:10.1007/BF02294770

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the beck depression inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*, 255–270. doi:10.1037/1040-3590.6.3.255

Sijtsma, K. (2005). Nonparametric item response theory models. In *K* (pp. 875–882). Encyclopedia of Social Measurement . New York: Elsevier.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric irt analysis of quality-of-life scales and its application to the world health organization quality-of-life scale (whoqol-bref). *Quality of Life Research*, *17*, 275–290. doi:10.1007/s11136-007-9281-6

Sijtsma, K., & Meijer, R. R. (2006). Nonparametric item response theory and special topics. *Handbook of Statistics*, *26*, 719–746. doi:10.1016/S0169-7161(06)26022-X

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. doi:10 . 4135 / 9781412984676

Stout, W. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement*, *25*, 300–306. doi:10.1177/01466210122032109

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educational and Psychological Measurement*, *71*, 834–848. doi:10 . 1177 / 0013164410393238

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, *21*, 22–40. doi:10.1080/08957340701796464

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79. doi:10.1037/1082-989X.12.1.58

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249. doi:BF02294536

Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.

Appendix: Kernel Smoothing IRT – Description of the Procedure

In parametric IRT models, a specific function is used for modeling of the relation between the latent trait and the probability of a correct/positively keyed item response, such as logistic or normal ogive function. For example, in the case of a two-parameter logistic IRT model,

$$P\left(y_{i,j}=1|\theta_i,a_j,b_j\right)=\frac{1}{1+exp^{-Da_j(\theta_i-b_j)}} \tag{1}$$

where the left side of the equation describes the conditional probability that examinee i's response y to a dichotomous item j is correct, θ_i is the latent dimension, α_j is the item discrimination and b_j is the item difficulty. The constant D is a scaling factor; when $D=1.7$, values of P for the two-parameter normal ogive and the two-parameter logistic models are very similar.

In KSIRT, nonparameteric estimation of OCC is performed without assuming any mathematical form for the OCC, based on kernel smoothing procedure. As described in Ramsay (1991, 2000), the main steps in KSIRT involve estimating the rank for each respondent based on their total scores (other estimate of ability can also be used); replacing the ranks by the quantiles of a certain distribution (commonly standard normal distribution); sorting respondents' response patterns by estimated ability rankings; and estimating the relation between item response and the latent dimension by kernel smoothing procedure at certain selected points (evaluation points). In the smoothing procedure, estimation is based on local averaging instead of using all the data points – the responses close to targeted scores are taken and approximation to regression curve is produced. Kernel is a weighting function, which assigns weights to the scores, based on their distance from the targeted score. There are different functions that can be chosen, with Gaussian, uniform, and quadratic as the most commonly used. In addition to the choice of a kernel function, the estimation process requires the choice of bandwidth (h). Bandwidth is a scaling factor that controls how wide the density estimate is spread around the point, that is, it controls the smoothness/roughness of the density estimate. Its inappropriate selection can lead to over- or under-smoothing of the curve. Selection of bandwidth assumes a trade-off between estimation bias and variance – larger bandwidth for example leads to smaller variance but larger bias.

Nonparametric OCCs are estimated by going through the above described steps. The independent variable is latent trait, the dependent variable is the probability of choosing the option m for item i, with the actual choices summarized

numerically by defining the variable y_{ima} (indicator variable) with the values 1 or 0 (when examinee a chooses this option or not). The probability function $P_{im}(\theta)$ is estimated by smoothing the relationship between these 0/1values and the examinee abilities by local averaging, in which for any proficiency or trait level θ the probability of choice $P_{im}(\theta)$ at that level is a weighted average of the values of y_{ima} for respondents with proficiency or trait levels close to θ

$$P_{im}\left(\theta_q\right) = \sum_{a=1}^{N} w_{aq} y_{ima} \tag{2}$$

where w_{aq} is a weight assigned to each examinee at each evaluation point q

$$w_{aq} = \frac{K\left[\left(\theta_a - \theta_q\right)/h\right]}{\sum_{b=1}^{N} K[(\theta_b - \theta_q)/h]} \tag{3}$$

with kernel function K and bandwidth parameter h. Detailed technical description of KSIRT and nonparametric regression technique can be found in Eubank (1988); Härdle (1990); and Ramsay (1991).

Performed Analyses

In this demonstration, a recently released R package KernSmoothIRT (Mazza et al., 2014) has been used. The authors state that all analyses that could be performed in the traditionally used TestGraph software (Ramsay, 2000) can be performed in KernSmoothIRT. General instructions for using R for data analysis are not included here, as many resources about the use of R are available (e.g., Revelle, 2017). R version 4.0.2 was used, and KernSmoothIRT version 6.4.

The demonstration data can be found in the file `ksirtfile.txt`, at https://osf.io/htcxe/quickfiles.

After the file is downloaded and saved, the data can be opened in R from the location where it was saved with:

```
ksirtfile <- read.delim("...file␣location/ksirtfile.txt", header=FALSE)
```

The KSIRT analyses were performed as follows, with first loading the R package `KernSmoothIRT`:

```
library(KernSmoothIRT)
```

Figure 1. OCC curves in Figure 1 were plotted by applying the `ksIRT` function and the `plot` function:

```
ksmooth<-ksIRT(responses = ksirtfile [ ,1:8], 1, 1, kernel = c("gaussian"), miss =
    c("option"), NAweight = 0, bandwidth = c("Silverman"), RankFun = "sum",
    thetadist = list("norm",0,1), groups = FALSE)

#individual graphs, for each item, are produced  with this command - press enter to
     see the next graph
plot(ksmooth, plottype="OCC", items = c(1:8),  axistype = "scores")
```

Figure 2. In the `plot` function, the chosen plot type was `EIS`:

```
plot(ksmooth, plottype="EIS", items= c(1:8))
```

Figure 3. For Figure 3b (ETS as a function of the quantiles of the standard normal distribution), plot type "expected" was selected:

```
plot(ksmooth, plottype="expected", items= c(1:8))
```

Figure 4. OCC curves in Figure 4 were plotted by applying `ksIRT` function and the `plot` function. Gender variable is included as "groups" in `ksIRT` function:

```
gender<-as.character (ksirtfile$V9)
DIF<-ksIRT(responses=ksirtfile[ ,1:8], 1, 1, groups = gender)
plot(DIF, plottype="OCCDIF",  items= c(1:8))
```

Figure 5. For Figure 5b (Q-Q plot pairwise expected scores for males and females), plot type "expectedDIF" was selected:

```
plot(DIF, plottype="expectedDIF", lwd=2, items= c(1:8))
```

Figure 3a and 5a - Density Plots. Figure 3a (Kernel density estimate of the distribution of the total score) was plotted with:

```
attach (ksirtfile)
total<-V1+V2+V3+V4+V5+V6+V7+V8
density <- density(total, bw=0.45)
plot(density, main = "Score␣Distibution", xlab = "␣", lwd=2)
```

Figure 5a (Kernel density estimate of the distribution of the total score for males and females) was plotted with:

```
library (lattice)
densityplot(~ total, group = gender, data = ksirtfile, bw=0.45,  plot.points=FALSE,
     xlab = "", main = "Score␣Distribution", lwd=2)
```

Citation

Rajlic, G. (2020). Visualizing items and measures: An overview and demonstration of the kernel smoothing item response theory technique. *The Quantitative Methods for Psychology*, *16*(4), 363–375. doi:10.20982/tqmp.16.4.p363

Received: 14/05/2020 ∼ Accepted: 04/08/2020