Przegląd Statystyczny. Statistical Review, 2020, vol. 67, 1, 51–76 DOI: 10.5604/01.3001.0014.1785

A proposal for perception measurement on a linguistic scale coded with unconventional fuzzy numbers

Marta Dziechciarz-Duda[a]

Abstract. The aim of this paper is to formulate a new proposal for perception measurement on a linguistic scale coded with fuzzy numbers. Additionally, an attempt is made to show the assessment process of the adequacy of a linguistic scale. The basis for the proposal is the discussion of issues related to the ambiguity of the results of measurements made by means of a subjective type of measurement scales. The proposed assessment technique is relevant when the results of the measurement based on a linguistic scale are coded with numerical equivalents in the form of e.g. unconventional fuzzy numbers.

The issue the subjective perception of the products' quality illustrates the objectivity level of measurement results. Subjective perception is measured with a specially designed IT tool allowing the respondent to determine all the characteristics of the resulting fuzzy numbers. The scale adequacy assessment tool is based on the Item Response Theory, and particulary so on the model devised by Georg Rasch.

The measurement of socio-economic phenomena, including material and subjective wellbeing of households, the quality of households' durable goods, and the assessment of the quality of goods available on the market requires special tools. It seems that one of the most useful and powerful tools for the measurement of socio-economic phenomena is a linguistic scale. The problematic issue in the analysis presented in the paper is coding verbal terms with their numerical equivalents.

Keywords: measurement, measurement scale, measurement scale adequacy, Item Response Theory, the Rasch model

JEL: C12, C52, C81, C82, C83

1. The introduction and motivation

The problem examined in the paper is the measurement of households' subjective perception of wellbeing. It is challenging to measure concepts such as wellbeing and its perception directly with a numerical scale, as both these notions are of qualitative nature (Tov and Diener, 2009). The existing measurement tools, such as ones based on the Likert scale, the 'divide 100 points' scale, semantic scales or benefit structure analysis do not address the heterogeneity of perception and subjectivity precisely enough. In other words, the existing measurement tools fail to recognise the variety and heterogeneity of respondents' statements (Walesiak and Gatnar, 2009). The research presented in this paper is aimed at designing new, improved measurement techniques appropriate for this type of socio-economic phenomena. The problem of household wellbeing belongs to a wider class of socioeconomic problems, where a subjective perception is a decisive factor for the measurement

[a] Wroclaw University of Economics and Business, Department of Econometrics and Computer Science, e-mail: marta.dziechciarz@ue.wroc.pl, ORCID: https://orcid.org/0000-0001-5038-9868.

results and, as such, requires a special means of measurement (Michalos, 2014; Fattore et al., 2011). One of the most important areas of subjective perception measurement is the quality and dignity of life.[1]

The importance of the results of the measurement of subjective consumer preferences for business could be described by e.g. marketing managers or decision-makers, who have benefitted from the valuable information yielded by this kind of measurement. The issue of subjectivity of consumers' perception is equally important with regard to new, innovative durable goods. Dynamic technological changes nowadays make it impossible for consumers to form informed opinions about such products based on objective technical and technological facts, so instead, they form their judgements on the basis of subjective impressions.

The definition of household wellbeing distinguishes among several approaches (ONS, 2019a, 2019b; Saisana, 2014).[2] The phenomenon of the households' subjective perception of wellbeing requires extensive discussion, which was initiated in the work by Diener et al. (2018). At this point it should be clarified, though, that this paper will not deal with the theoretical concept of the subjective perception of wellbeing, but will focus on the early stages of introducing improved measurement techniques of the subjective perception of qualitative phenomena, as seen from the methodological point of view.

Since it is difficult to quantify perceptions on a metric scale, the researcher may request the respondents to use verbal, linguistic phrases to describe their perception of the object of interest. Linguistic variables seem to be one of the most promising techniques for measuring socioeconomic phenomena. A linguistic variable is one whose values are presented in the form of verbal categories, to which, in turn, numerical codes are assigned. In the measurement practice, linguistic terms are used to measure the status of the selected socio-economic phenomena, which include the subjective perception of welfare, the subjective assessment of the quality of a household's durable goods, etc. Wherever subjectivity is involved, linguistic variables are convenient and intuitive means of assessing perceptions or preferences, since their values are defined as verbal categories. Therefore, linguistic variables make it possible to quantify the criteria for phenomena which by nature are categorical and potentially perceived differently by every respondent (*FisPro…*, 2018, p. 45–48). The usability of linguistic variables in the studies of socio-economic phenomena is moreover indicated in literature, for example in the work by Schnorr-Bäcker (2018).

[1] OECD Better Life Index, OECD, Paris, www.oecdbetterlifeindex.org (accessed 20.11.2019).

[2] Quality of Life Research, selected issues of an Official Journal of the International Society of Quality of Life Research (*International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*).

As was mentioned before, it is difficult to interpret results obtained by means of a subjective type of measurement scales unequivocally, and this is how the method of quantifying linguistic expressions became the subject of this paper. It is important to remember here that unless fuzzy numbers are applied to encoding verbal statements, it is difficult to formulate clear recommendations on how to determine the domain of fuzzy numbers for a verbal variable. The author's experience relating to the measurement of socio-economic phenomena (including the material wealth of households determined by their possession of durable goods) leads to the conclusion that coding linguistic variables, i.e. verbal statements, into numerical equivalents is most effective when performed with the use of the so-called unconventional fuzzy numbers, i.e. numbers which have an uneven length, are unbalanced and are of an overlapping shape (Arguelles Mendez, 2016; Roubens and Vincke, 1988).

The application of linguistic variables is beneficial for the respondent, but the researchers are left with a difficult task of coding verbal statements into numerical equivalents adequately and choosing the analytical techniques and the base for inference. Using fuzzy numbers for this purpose is one of the possible approaches (Zalnezhad and Sarhan, 2014; Kacprzyk and Roubens, 1988; Zadeh, 1975). The aim of this paper is to estimate the degree of adequacy and precision of linguistic variables used as measurement tools for subjective assessment. In this context, the latent trait[3] models seem to be promising instruments of the assessment of scale adequacy. They are designed to measure the underlying ability or trait, which the test result indicates, rather than measuring the performance *per se*. Another argument in favour of latent test models is that the structure of the test is sample-free. The results are independent of the measurement scheme which generated them. The method of latent trait models was developed in the 1950s, but due to the lack of specialised computer software, it could not be used in empirical research and thus for a long time it remained a theoretical concept with no practical applicability.

The Item Response Theory Models seem to be an adequate tool for the assessment of the relevance and precision of measurement scales based on linguistic variables. Measuring and assessing the scale adequacy helps to improve the coding quality in the process of replacing verbal statements with responses in the form of unconventional fuzzy numbers. The Item Response Theory[4] (IRT) is a theoretical system of models, including probabilistic ones, applicable to the analyses and evaluation of measurement scales. The technique is associated with the name of one of the authors, Georg Rasch. The scaling used in the IRT models assumes that anyone who can re-

[3] Trait is a distinguishing feature of a person's character. Often, in literature on the subject, a trait is called respondent characteristic or respondent ability.

[4] The adopted convention is that names which can be used in an abbreviated form are written with capital initials, e.g. Item Response Theory could also be referred to as IRT.

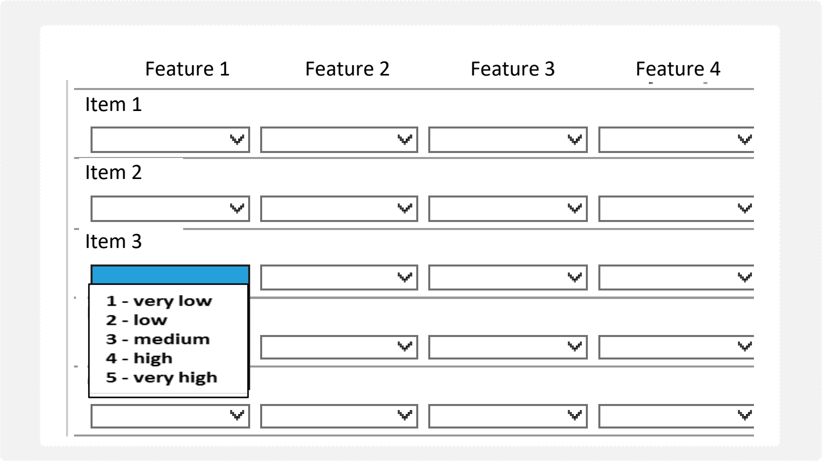
spond to statements of high difficulty will be able to respond to statements of low difficulty, too. The Item Response Theory models for the assessment of test items and questions belong to a group of models gradually developing in social research and applicable to the following areas: psychology, education (for example in the PISA study), medicine and marketing. The family of IRT models is rooted in theories devised by L. Guttman and R. Mokken, who introduced non-parametric probabilisation of the Guttman scalogram (Guttman, 1944; Guttman et al., 1950; Hofmann, 1979; Abdi, 2010; Mokken, 1971; Sijtsma and Ark, 2017; Ark, 2012; Wind, 2017; Watson et al., 2018).

In order to measure the subjective perception of socioeconomic phenomena effectively, the best solution seems to be, as mentioned before, the application of a linguistic scale with verbal categories used for determining the assessment results of a group of respondents. It has also been noted that numerous characteristics of socio-economic phenomena are inherently qualitative, therefore conventional, quanti-tative measurement tools fail to fully overcome problems which often occur in the process of measuring the perceptions or attitudes of respondents. The aforementioned problems lie in the fact that the researcher attempts to quantify characteristics which are either immeasurable (on metric scales) or hidden. For this reason, an alternative approach needs to be applied, involving the use of verbal, linguistic phrases to describe such characteristics as attitudes towards or perception of the phenomenon of interest. Verbal, linguistic phrases that attempt to capture and explain the differences in the individual respondent's assessments of those phenomena constitute the measuring technique recommended in socio-economic analyses (European Commission, 2017; Zamri and Abdullah, 2014).

2. Conceptual framework. Measurement method. Linguistic form of characteristics' level determination

Although the advantages of using linguistic variables, which are intuitive and convenient for respondents to express their judgements opinions or preferences, have already been discussed, they cannot be overestimated. A linguistic variable is a form of a characteristic whose values are determined using a verbal category. The linguistic form of a characteristic is referred to as a linguistic feature. It may be used by the respondent for the description of the subjectively perceived level of the measured characteristic of a socio-economic phenomenon. Linguistic features have values defined as verbal categories. But, as was mentioned before, using linguistic features poses serious problems for the researcher, who has to face the challenging task of adequately encoding verbal statements into numerical values. Along with the advantages of linguistic variables comes a basic conceptual problem – phenomena are inherently descriptive and as such, having the potential of being understood differently by different respondents. One of the possible solutions to this problem is to employ a linguistic description, which might further be translated, i.e. coded, into some form of numerical values. The algorithm of the procedure involves the use of linguistic variables to determine a respondent's assessment of a given phenomenon by indicating one of the verbal levels of the linguistic variable. Subsequently, linguistic variable levels are assigned to their numerical equivalents, i.e. coded usually into some forms of fuzzy numbers. Both steps are performed by the respondent. In the first step, respondents select verbal categories.

Figure 1. Respondents' choice of verbal categories included in the assessed items

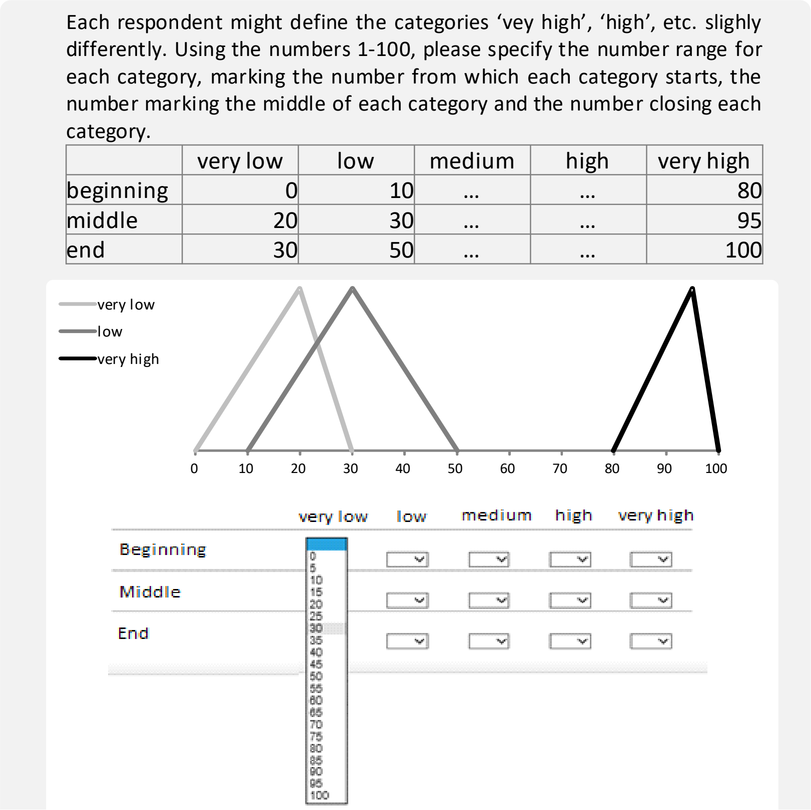


	Feature 1	Feature 2	Feature 3	Feature 4
Item 1				
Item 2				
Item 3				

Source: author's work.

In the second step, respondents perform the coding of verbal categories with numerical values. Figure 1 presents an example of a computer screen view for respondents with instructions on how to define the lower, medium, and upper value. The respondent may see the immediate graphical illustration of coding verbal categories in the form of triangles. The outcome of step 1 and 2 of the measurement procedure applied to all respondents is obtaining the measurement results in the form of fuzzy numbers. Such numbers take an unconventional form. Figure 2 illustrates the possible shapes of triangle fuzzy numbers attached to categories (very low, low, very high). The numerical values are given by only one respondent who was selected for illustrative purposes.

Figure 2. Respondents' choices. Coding verbal categories with numerical values



Each respondent might define the categories 'vey high', 'high', etc. slighly differently. Using the numbers 1-100, please specify the number range for each category, marking the number from which each category starts, the number marking the middle of each category and the number closing each category.

	very low	low	medium	high	very high
beginning	0	10	...	...	80
middle	20	30	...	...	95
end	30	50	...	...	100

	very low	low	medium	high	very high
Beginning					
Middle					
End					

Source: author's work.

The main concern of a researcher using this measurement procedure is to guarantee a satisfactory level of adequacy, accuracy, and precision of the measurement. It is the researcher's responsibility to ensure that the variable used to measure attitudes and perceptions was described by an adequate measurement scale. The next issue is to identify the quantitative equivalents for verbal expressions used to mark various verbal categories of a natural language used as levels of the linguistic variable. When linguistic variables and verbal categories are coherent, it is possible to use the idea of Georg Rasch to enhance the uniformity in the interpretation of the assessments. Testing the adequacy of the measurement scale is vital to ensure that the measurement results are satisfactorily accurate.

3. An outline of the Item Response Theory. Model formulation

The idea of item analysis is based on the assumption that there is a hierarchy describing the quality of survey questionnaires. Discussing the issue, DePaoli et al. (2018, pp. 1299–1300) said: 'From a survey-development perspective, it is important to thoroughly examine the psychometric properties of any survey before finalizing the measure for broad use. [...] there are other techniques based on the item response theory (IRT) framework that provide a more detailed assessment of the survey items'. In this context, placing the Rasch model on the broader outline of the Item Response Theory seems worthwhile (Royal et al., 2010; Zhu, 2002). In its mathematical concept, the Rasch model is a special case of the Item Response Theory, namely a one-parameter IRT model,[5] called the one-parameter linear model (1PL). In the literature on the subject, the IRT concept is compared with the Classical Test Theory (CTT). The extensive comparison of this kind may be found in the seminal work by R. Jabrailov et al. The authors state that 'The crucial difference between CTT and IRT is that in CTT the cutoffs are based on the distribution of the sum scores X, whereas in IRT they are based on the probability distribution' (Jabrayilov et al., 2016, p. 560). In the IRT context, the cutoff would be a certain quantile, usually a high percentile of the probability distribution in a functional population. Specialised websites are a comprehensive source of publications on the theory and applications of Rasch-type models (Jumailiyah, 2017).[6] An excellent comparison of the theoretical foundations of CTT and IRT may be found in Chapter 2 of the fundamental monography by DeMars (2018). The author discusses the accepted assumptions and formulates the specification of base model types and rules governing the process of designing the scale levels for items, along with practical issues, including reliability, required sample size, etc. The analytical review of cognitive and application aspects of CTT and IRT is presented by several authors, including Kong (2018) and Raykov and Marcoulides (2016). In their comprehensive analysis, Jabrayilov et al. (2016) showed the empirical outcome of the comparisons of the results of applications.

Due to readily available statistical packages, CTT are the easiest and most widely used techniques in the field of statistical analysis. The difference between IRT and classical analyses is that classical testing is usually performed on an entire set, whereas IRT more often concentrates on one item. What is more, the application of CTT is limited solely to the analysed population; sample and inference cannot be

[5] For simplicity, wherever possible and convenient, the Rasch model will be called the IRT model.

[6] Institute for Objective Measurement, https://www.rasch.org (accessed 20.11.2019). *Journal of Applied Measurement*, selected issues, http://jampress.org (accessed 20.11.2019). Rasch Measurement Transactions Contents, Archives of the Rasch Measurement, selected issues, https://www.rasch.org/rmt/contents.htm (accessed 20.11.2019).

extended onto items belonging to another sample,[7] although CCT can also generate item statistics. The question of the advantage of one approach over the other is subject to discussion. R. Jabrailov et al. state that 'The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations' (Jabrayilov et al., 2016, p. 559). Other authors claim that the application of CTT or IRT depends on the nature of testing situations, as each approach has its specific advantages and disadvantages. A full list of key advantages of IRT over CTT is given by Prieler (2007) (see Table 1).

Table 1. Key advantages of IRT over CTT for the analysis of change

CTT	IRT
The relation between the score and the ability level is based on the overall score across items.	A direct relationship is established between the ability level and the parameters of individual items (such as the difficulty of the item and discriminative power at different points in the distribution).
Emerging factors are seen as 'primary' influences on the test performance, with individual items being affected in different ways by other factors.	Emerging factors are less influenced by secondary factors, as much attention has been devoted to the issue of item homogeneity.
'Bad' items reduce the predictive power.	'Bad' items are eliminated.
Level of ability is defined in relation to a particular sample.	Level of ability can be defined independently of any sample.
Correlation is used to compare performance on repeated test occasions, which obscures the analysis.	No need to use correlation, so disadvantages are removed.
It is not possible to measure the significance of change at the individual level.	The significance of change at the individual level can be objectively measured.

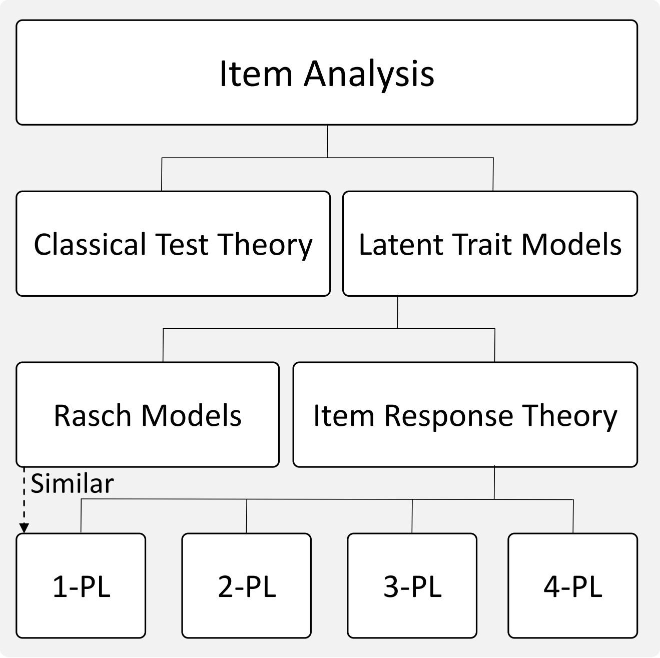
Source: author's work based on Prieler (2007, p. 701).

Jabrailov et al. (2016, p. 559) assert that researchers are able to see the possible advantages of using IRT over CTT. According to the authors, in the situation where tests consist of at least 20 items, the comparison of the CTT and IRT methods with regard to Type I error and detection rates showed that IRT is indeed superior to CTT in individual change detection. On the other hand, CTT appeared to be more effective at correctly detecting the change in individuals in shorter tests. Similar results were reported by Magno (2009). The popularity of the Item Response Theory results from social testing programs, conducted frequently and on a large scale, where IRT is referred to as modern psychometrics. This is a consequence of large-scale education assessment (e.g. PISA) or professional market testing (Edelen and Reeve, 2007). It could be said, in general terms, that IRT has many advantages over CTT that have brought it into more frequent use (Hambleton and Swaminathan, 1991c). It also seems that IRT has almost completely replaced CTT as a method of choice in some areas of application.

[7] Descriptive IRT vs. Prescriptive Rasch, https://www.rasch.org/rmt/rmt51f.htm (accessed 20.11.2019).

The question of the quality of measurement scales is connected with the issue of how adequate the survey questions were for respondents and to what extent they measured the ability of respondents to provide correct answers. The general framework of tests, questionnaires, and surveys makes it possible to reuse items such as e.g. questions or questionnaires, and thus they can appear repeatedly in several such structures. This is because their quality has already been verified, i.e. it is known how the questions are going to perform. In other words, IRT enables creating a reservoir of questions of foreseeable performance. Such a reservoir may constitute a kind of databank of questions and questionnaires (Combrinck, 2018; Yau and Yao, 2011; Linacre, 2002). Figure 3 provides an overview of the Item Response Theory with an indicated relation to the Classical Test Theory framework. The position of the Rasch model within the Item Response Theory models is also specified.

Figure 3. Classification of Item Response Theory models



Source: author's work based on Hambleton and Swaminathan (1991a).

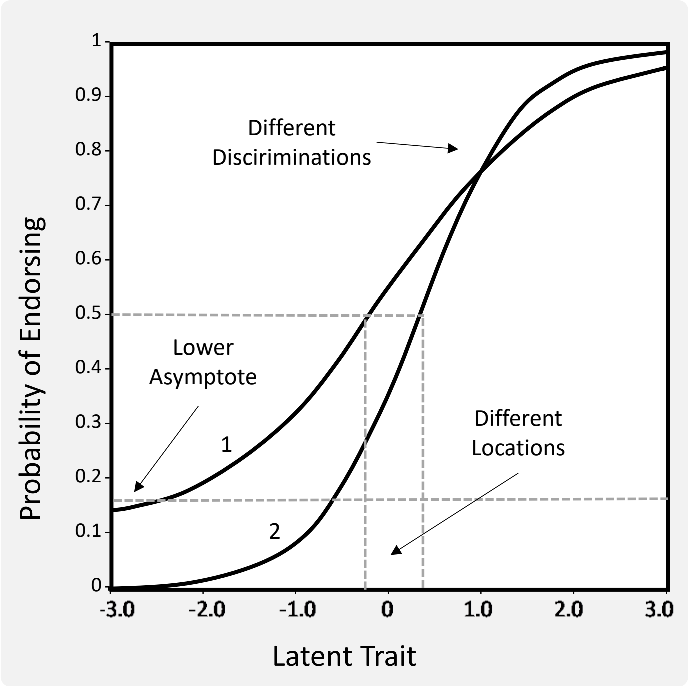
The Item Response Theory framework consists of three basic components:

- Item Response Function (IRF) – the function that relates the value of the latent characteristic (trait) to the probability of endorsing an item. The basis of the definition of the IRF is the concept of the latent variable defined as individual differences in reaction (assessment) to a construct (item). The IRF expresses the relationship between a latent variable as defined above and the probability of endorsing an item. The concept of the IRF is used for modelling the relationship between the respondent trait level, the item properties, and the probability of endorsing the item;

- Item Information Function (IIF), which is considered as the indicator of the quality of an item, i.e. the item's ability to differentiate among respondents;
- invariance; provided that invariance is sustained, it is possible to estimate the item parameters from any position on the item response curve. Similarly, it is possible to estimate the parameters of an item from any group of respondents who have answered the item.

Item Characteristic Curves (ICC) is a very useful tool for representing results in a graphical form. ICC is created by the conversion of IRF into graphical functions which represent the respondent's abilities. The values of the ICC function represent the probabilities of endorsing an item by the respondent. The role of the item discrimination parameter is to illustrate the steepness of the IRF for each location of the item, i.e. the strength of the relation between the item and the value of the latent characteristic (Figure 4). Here, the analogy between the latent trait and the loadings in factor analysis might be observed. Items with a high discrimination parameter value may appear as ones which are better at differentiating respondents around the location point. In other words, minor changes in the latent trait lead to significant changes in the probability value. The latter statement also applies to the opposite – items with a low value of discrimination parameter a may be viewed as ones which are not as effective in differentiating respondents around the location point. Item location parameter b is defined as the amount of the latent trait which covers at least half of the probability of endorsing the item. The rule is that the higher the respondent's trait level while attempting to endorse the item, the higher value parameter b has. A similarity to the Classical Test Theory may be observed, as it involves the same complex task to determine Z scores. Additionally, as in the case of Z scores in CTT, usually the numeric values of parameter b range from -3 to $+3$. Applying parameter c, called item parameter guessing, increases the probability that respondents with a very low trait level may still choose the correct answer. One may expect that respondents presenting low trait levels, yet with a good intuition (and selecting answers at random) may still stand a chance of endorsing an item. It frequently occurs when multiple-choice testing is involved. It is expected that the parameter value should not vary considerably from the number of reciprocal choices. Parameter d of the IRF, called the item parameters upper asymptote, assumes that the probability that respondents with extremely high abilities will answer correctly is less than one. In other words, even such respondents are not always certain to make the correct choice.

Figure 4. Item Response Theory framework



Source: author's work based on Hambleton and Swaminathan (1991a).

The comprehensive four-parameter (4PL) logistic model may be denoted in the form of the following formula:

$$P(X = 1|\theta, a, b, c, d) = c + (d - c)\frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \tag{1}$$

where:

θ – represents respondent trait level,

a – denotes the item's discrimination that determines the steepness of the IRF, alternatively called item parameters discrimination,

b – denotes the item difficulty that determines the location of the IRF, alternatively called item parameters location,

c – denotes a lower asymptote parameter for the IRF, alternatively called item parameters guessing; restricts the probability of endorsing the correct response for respondents with extremely low ability,

d – denotes an upper asymptote parameter for the IRF, alternatively called item parameters upper asymptote, which restricts the probability of endorsing the correct response for respondents with extremely high ability.

The left side of Equation (1) indicates the probability of responding correctly to a given item according to the key of answers. In calculations, the number of items and the number of respondents is denoted $i = 1, \dots , n, j = 1, \dots , N;$ respectively. The four-parameter logistic model (4PL) is the most complex and comprehensive logistic model of the Item Response Theory. A reformulation of Equation 1 to a simpler formula, the 3PL model, is possible by removing parameter d, and the 2PL and 1PL models could be obtained analogically – by removing parameters c and b, respectively.

As stated before, from the formal, mathematical point of view, the model proposed by Georg Rasch is identical to the basic Item Response Theory Model (1PL). What is different here is the approach of Rasch himself, who believed the model to be superior to data. Following his way of thinking, if some data does not fit the model, it should be discarded. Additionally, Rasch's specification does not allow the estimation of abilities for extreme items and persons. In principle, the Rasch model is designed for categorical data. The model's elegant mathematical form renders it suitable for extensions that allow greater flexibility in handling complex samples relating to collections of tasks representing different domains. Extensions of the Rasch model are enhanced by additional structural elements that account for differences among diverse populations or observed variables.

4. The assumptions of the Item Response Theory models

The Item Response Theory is a universal paradigm. The variants of models are designed to suit the specific qualities of any given population. A set of common assumptions constitutes the base for the specification, the assessment of applicability and the rules for the interpretation of results (NAP, 2017; Tinsley and Brown, 2000; Hambleton and Swaminathan, 1991b).

The Invariance assumption. Invariance is the position of the latent trait which may be estimated by any item with a known Item Response Function. Invariance means here that the item characteristics are independent of the population to which they are applied. The statement concerns the linear transformation of items. So, invariance means that regardless of which questions are being asked, the assessment of the level of respondent's abilities remains the same. In other words, the assessment of the level of respondents' abilities does not change when the questions do. On the other hand, item parameters are not determined by a particular group within the sample of respondents or inside their linear transformation. The property or assumption of invariance is crucial for socio-economic measurement. It makes it possible to:

- link scales that measure the same construct;
- implement computerised adaptive testing;
- compare respondents, also when they answered different items on the scale.

Unidimensionality assumption. It is assumed that in the conventional Item Response Theory models, considered to be one-dimensional, parameter theta characterises individual differences. As a consequence, the item covariance in the discussed model specification includes a single common factor, i.e. a latent trait or a latent feature, which is estimated by means of specialised factor analytic models for dichotomous items (Maydeu-Olivares et al., 2011; Kappenburg -ten Holt, 2014). There are also multidimensional IRT models, but they are not commonly used in applied research (Immekus et al., 2019; Hartig and Hoehler, 2009; Ackerman, 2005).

Local Independence Assumption. The Local Independence (LI) assumption indicates that item responses are uncorrelated, provided that control over the latent trait is established. The LI and unidimensionality are naturally related. The former is liable to violations which are called local dependencies. Even highly univocal scales can be susceptible to violations of local independence, which may occur, for example, due to item content dependence. Local Independence Assumption violations may lead to serious consequences (local dependencies), such as the following:

- they may distort values of item parameter estimates, which in practice means that item slopes are steeper than they really are;
- they may cause the scale to look more precise than it actually is;
- the occurrence of Local Dependence (LD) may lead to a false conclusion about the invalidity of the scale, which may essentially define or dominate the latent trait in a construct where a strong correlation between two or more items exists.

Therefore, the violations of local independence have to be addressed. H. Wainer and G. Kiely recommended forming testlets by combining locally dependent items for this purpose. A testlet is defined as an aggregation of items which are based on a single stimulus, such as, for example, a reading comprehension test. In this case, a testlet is a passage and the set of four to twelve items that accompany it (Wainer and Kiely, 1987; Sireci et al., 2005). Alternatively, LD may be addressed by removing one or more of the LD items from the scale in order to achieve local independence.

Qualitatively homogeneous population assumption. The key assumption of the Item Response Theory models states that the same IRF applies to all members of the respondent's population. The violation of the qualitatively homogeneous population assumption, called differential item functioning, means that a violation of the invariance property occurred. If an item has a different IRF for two or more groups, it may lead to false conclusions, e.g. for respondents who are equal in terms of the latent feature, different probabilities of the expected scores of endorsing an item could be estimated.

Monotonicity assumption. When specifying logistic IRT models, it is assumed that as the trait level increases, so does the probability of endorsing an item. In mathematical terms, models have the form of a monotonically increasing function. As a consequence, in the situation where this assumption is violated, applying the logistic form of the model to describe item response data becomes pointless.

5. Item Response Theory model – application

The concept of the Test Response Curve (TRC) is crucial for the interpretation of the results from the point of view of their applicability. Since Item Response Functions are additive, the researcher can combine items to create a Test Response Curve. TRC describes the latent trait's dependency on the number of considered items. An equally important analytical tool is the Item Information Function (IIF), where the item reliability is replaced by the item information. Each IRF can be transformed into the IIF. The values of IIF provide a precise representation of an item at each level of the latent trait. The information has the form of an index representing the item's ability to differentiate among individuals. The standard error of measurement, which is a variance of the latent trait level, may be interpreted in such a way that more information means less error, and vice versa. According to the standard error definition, the measurement error is expressed on the same metric scale as the latent trait level, so it can be used to build confidence intervals. The Item Information Function is crucial in the process of creating a quality description of the measurement scale. It is possible to extract several of its characteristics, including the following:

- a difficulty parameter, understood as the location of the highest information point;
- a discrimination parameter, understood as the height of the information;
- large discriminations, i.e. the high and narrow IIFs; a high level of precision is expressed by a narrow range;
- low discrimination, i.e. short and wide IIFs; a low level of precision is expressed by a broad range.

In the one-parameter logistic model, the discrimination parameter is fixed for all items, and, accordingly, all the Item Characteristic Curves corresponding to different items on the measurement scale are parallel along the ability scale.

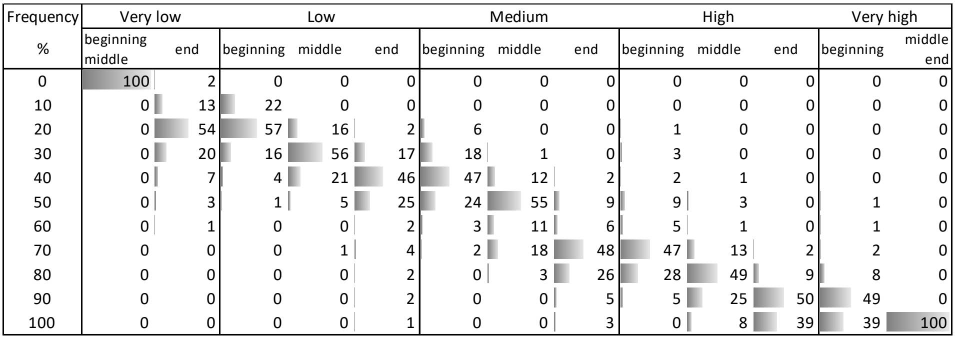
The Item Information Function values are the measurements of the amount of information provided by individual items. Those values may be calculated by multiplying the probability of endorsing a correct response by the probability of endorsing an incorrect answer.

Since the Item Information Functions are additive, the aggregate function may be understood as the Test Information Function (TIF). The TIF may be used for the assessment of the test as a whole, and in particular to identify parts of the characteristic range that are most precise and perform best. In Polish literature, the papers by Jefmański (2014) and Brzezińska (2016) constitute the first attempts to show the potential of this methodology.

6. Assessment of scale adequacy

The data which was used for the exercise that served as an example of the assessment of the scale adequacy within the Rasch theoretical framework was collected at the Wroclaw University of Economics, i.e. the university the author works at. The data was collected according to the procedure described in part three of this article and by means of the Computer Assisted Personal Interview survey (Lynn, 2019). Respondents were asked to specify their opinions concerning a set of innovative products (smartphones), using computer screens illustrated in Figures 1 and 2. The products were described by five characteristics and the overall assessment variable. Altogether, the sample consisted of over 450 sets of assessments submitted by the respondents. Since the group of respondents was selected using the convenience approach, the study should then be considered a pilot study whose aim was to verify the possibility of applying the proposed approach. The questionnaire covered the following issues: respondents' preferences as to the leading smartphone brands, the available smartphone applications, and the key characteristics of the devices. The measurement results were collected in the form of unconventional fuzzy numbers. Figure 5 illustrates the frequency (%) of chosen answers defining the beginning, middle, and upper limit of fuzzy numbers corresponding to the individual categories of the verbal grades.

Figure 5. Frequency (in %) of answers defining the beginning, middle, and upper limit of fuzzy numbers corresponding to individual categories of verbal grades

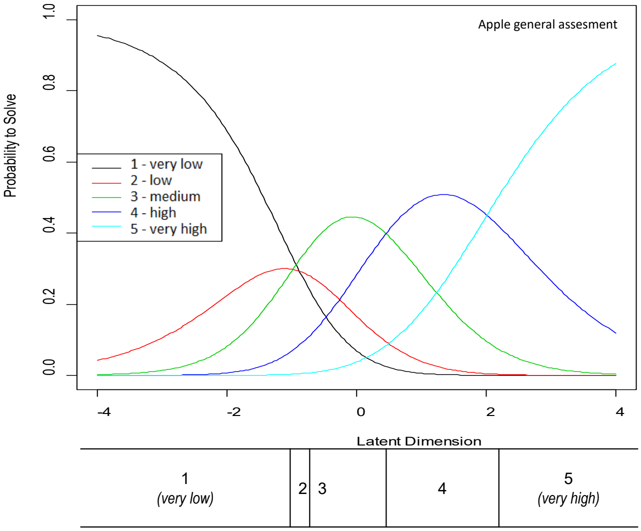


Frequency	Very low		Low			Medium			High			Very high	
%	beginning middle	end	beginning	middle	end	beginning	middle	end	beginning	middle	end	beginning	middle end
0	100	2	0	0	0	0	0	0	0	0	0	0	0
10	0	13	22	0	0	0	0	0	0	0	0	0	0
20	0	54	57	16	2	6	0	0	1	0	0	0	0
30	0	20	16	56	17	18	1	0	3	0	0	0	0
40	0	7	4	21	46	47	12	2	2	1	0	0	0
50	0	3	1	5	25	24	55	9	9	3	0	1	0
60	0	1	0	0	2	3	11	6	5	1	0	1	0
70	0	0	0	1	4	2	18	48	47	13	2	2	0
80	0	0	0	0	2	0	3	26	28	49	9	8	0
90	0	0	0	0	2	0	0	5	5	25	50	49	0
100	0	0	0	0	1	0	0	3	0	8	39	39	100

Source: author's work.

Respondents defined numerical values of the verbal categories with various forms of triangular fuzzy numbers. Some respondents attempted not to overlap, some tried to keep the equal length, while some others to cover the full range of possible values (from 0 to 100). A very useful tool for graphical representation of results is the Item Characteristic Curves. The ICC is the result of the IRF conversion which takes the form of graphical functions representing the respondent's ability.

Figure 6. Characteristic curves for subscale items (Item Characteristic Curves)



Source: author's work in R (eRm package).

The values of the ICC function are the probabilities of endorsing the item by the respondent. The general assessment of Apple smartphone is shown in Figure 6 to illustrate the location of borders between categories. Respondents are very far from a uniform distribution in their statements; the widest range is attributed to categories very low and very high, while the category low has a very narrow range attributed by respondents, which came as a surprise. On the other hand, there were some respondents who defined very narrow ranges of codes for their verbal categories, but also approximately a third of them coded their categories with wide and frequently overlapping ranges. Subsequent illustrations for the remaining characteristics shown in Figure 7 confirm that it was worthwhile to allow unconventional fuzziness for characteristics assessment. The uneven lengths of numerical codes attached to individual verbal categories prove that respondents attach various meanings and diverse, hidden and latent values connected to their subjective perception of product features. The interpretation of the assessments is strongly related to the shape of the resulting triangle. Narrow triangles signify strong opinions anchored in a well-established view resulting in the knowledge of the rules for interpreting attribute values. Wide triangles, on the other hand, signify the lack of strong opinions, no solid views and the lack of determination in formulating opinions.

In addition, it can be inferred that these respondents do not have proper knowledge about the products studied, or their general knowledge is poor. This leads to a specific interpretation of product features, i.e. the subjective perception of the product properties. As a result, formulated assessments of the subjective perception of individual product properties take the form of triangles with very broad foundations.

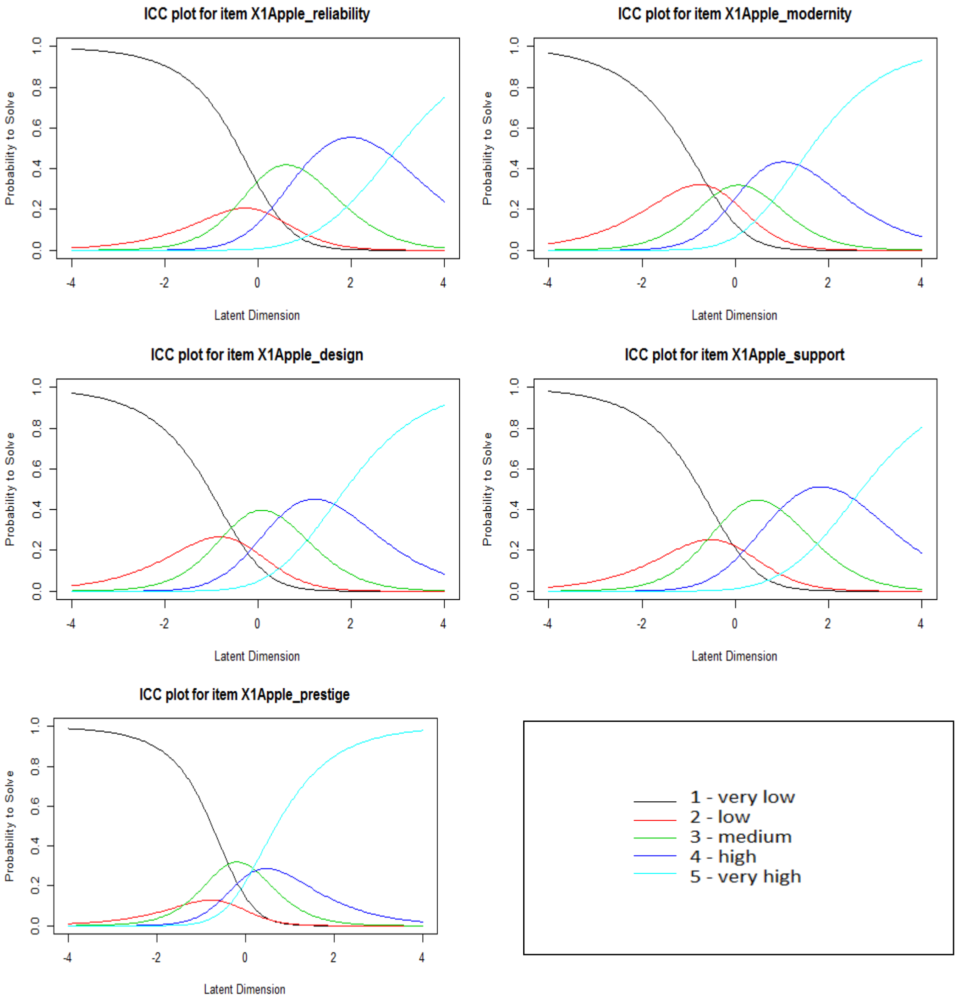
A response to the latter observation is the probabilistic test theory, which examines the probability which may be attached to the respondents' possible answers to a given scale item. Scale items, called statements, are a function of a hidden variable that specifies the level of the respondent's ability to measure the socio-economic phenomenon properly, understood as the ability to give a true answer to the scale item. On the other hand, the level of difficulty of test statements should be assessed. Both tasks may be done using the Rasch specifications.

The Item Information Function curve indicates the quality of an item, i.e. the item's ability to differentiate among respondents (Figure 8). The definition of the Response Function is based on a concept of the latent variable defined as individual differences in reaction, manifested in the assessment of a construct, sometimes referred to as an item. The IRF characterises the relation between such latent variable and the probability of endorsing an item.

As was mentioned before, Item Information Functions are additive and the aggregate function is called Test Information Function. As shown in Figure 9, TIF may be used as a complete assessment of a test. It is also helpful while identifying parts of the characteristic range that are most precise and perform best. Additionally, it might be considered an indicator of item quality, i.e. the item's ability to differentiate among respondents. Linguistic expressions may be coded as fuzzy triangular numbers by means of a partial credit model framework. This kind of model belongs to the family of models used for the theory of response to scale items. The ranges of the domains may be determined on the basis of the intersection points of the characteristic curves of adjacent categories (Linacre, 2000, 2002). The most recent results may be found in the summary provided on a specialised website.[8]

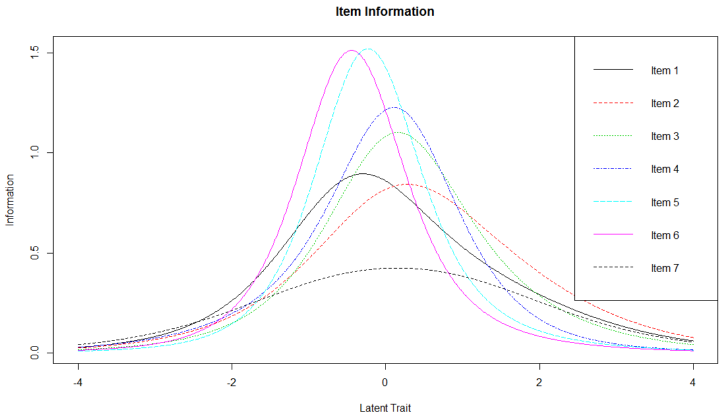
[8] Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM), https://www.rasch.org/rmt/rmt1231.htm (accessed 20.11.2019).

Figure 7. Characteristic curves for subscale items used for Apple assessment



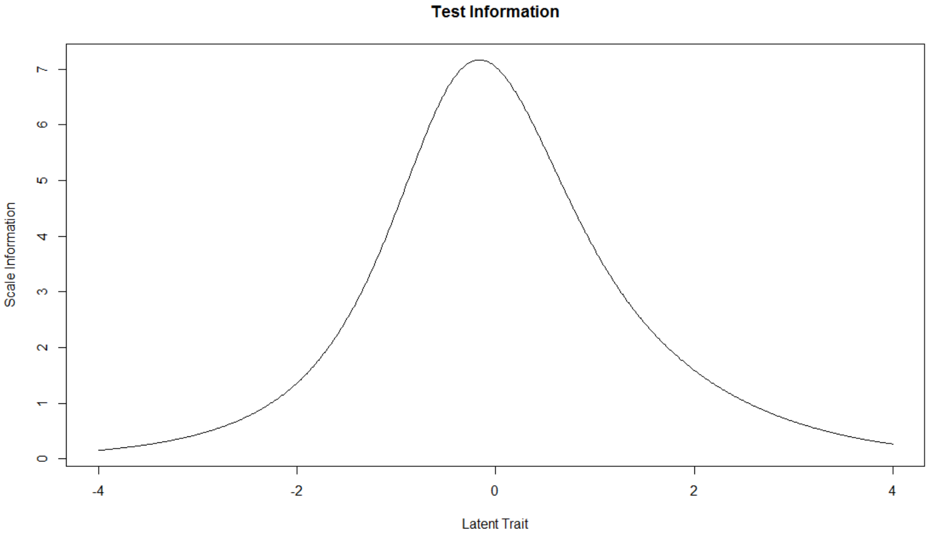
Source: author's work in R (eRm package).

Figure 8. Item Information Curves



Source: author's work in R (eRm package).

Figure 9. Aggregate function understood as Test Information Function



Source: author's work in R (eRm package).

Table 2 introduces formulas that enable the establishment of the parameters of triangular fuzzy numbers for each of the categories outlined within the i-th item. The example presents a rating scale with five verbal categories on the ordinal scale: very low (VL), low (L), medium (M), high (H), and very high (VH).

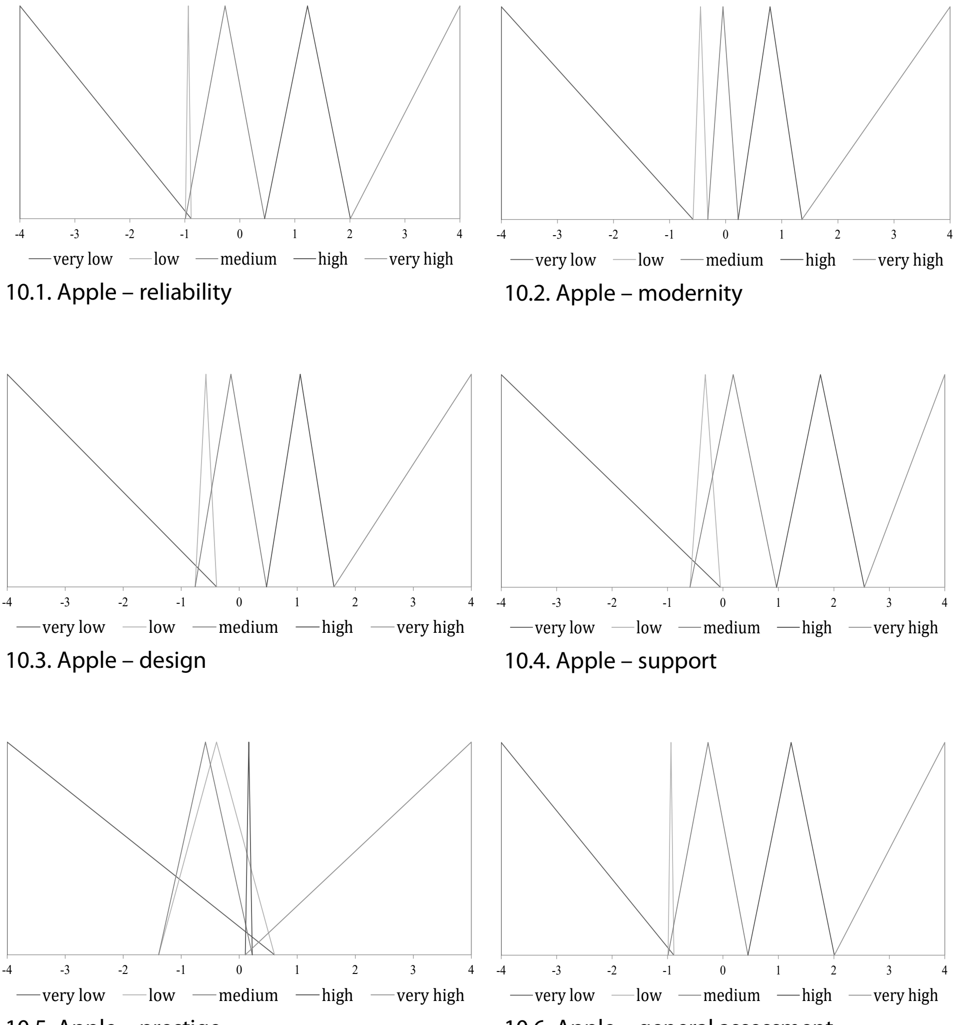
Table 2. Determination of the values of parameters of triangular fuzzy numbers for verbal categories

Parameters of fuzzy numbers	Categories				
	1 – Very low	2	3	4	5 – Very high
α_1	-4	τ_{i1}	τ_{i2}	τ_{i3}	τ_{i4}
α_2	-4	$\frac{\tau_{i1}+\tau_{i2}}{2}$	$\frac{\tau_{i2}+\tau_{i3}}{2}$	$\frac{\tau_{i3}+\tau_{i4}}{2}$	4
α_3	τ_{i1}	τ_{i2}	τ_{i3}	τ_{i4}	4

Source: author's work based on Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM), https://www.rasch.org/rmt/rmt1231.htm (accessed 20.11.2019); Linacre (2000, 2002).

It should be stressed that the category represented by the term *low* is hardly ever chosen, which seen from the methodical point of view might lead to the conclusion that this category could be removed from the scale, as according to the respondents' choices, it hardly describes the considered phenomenon at all. On the other hand, however, it might be that most respondents evaluate the product characteristic positively, whereas a negative value is chosen only by those respondents who do not accept the brand at all (for example see Figure 10).

Figure 10. A graphical form of triangular fuzzy numbers assigned to verbal categories for the Apple brand



10.1. Apple – reliability

10.2. Apple – modernity

10.3. Apple – design

10.4. Apple – support

10.5. Apple – prestige

10.6. Apple – general assessment

Source: author's work.

7. Concluding remarks

In conclusion, it may be stated that the measurement of socio-economic phenomena, including the subjective perception of the material wellbeing of households, the quality of durable goods in households, and the assessment of the quality of goods available to the members of the household requires special tools. It has been proven that one of the most useful and powerful of such tools is a linguistic scale. Specialised procedures are necessary to code verbal terms with numerical equivalents. The author suggests the use of unconventional fuzzy numbers for this purpose.

The new proposal on how to perform the assessment and measurement of the scale adequacy proved to be useful and effective. The idea of the discussed assessment technique becomes relevant when the measurement results of a linguistic scale are coded with numerical equivalents. The author is interested in increasing the objectivity of the results of the measurement of households' subjective wellbeing as well as the subjective perception of households' endowment with durables. The process of testing the author's theory included the measurement of the subjective perception of the socio-economic phenomena on a linguistic scale. The respondents coded their own subjective perceptions with fuzzy numbers, usually with unconventional forms of fuzziness. This confirmed the author's supposition of the diversity of individual perceptions and assessment of phenomena. The core of the author's interest is focused on the use of unconventional fuzzy numbers.

As Figure 6 indicates, it is possible to establish numerical delimitation points between verbal categories. The technique proves useful in the design of survey questionnaires. The framework for the assessment of scale adequacy is provided by the Item Response Theory. The author tested the usefulness of the one-parameter variant of the ITR, often called the Rasch model. That study demonstrated that the interpretation of assessments can be strongly related to the shape of the resulting triangles. Hence, it is advisable, and sometimes necessary, to analyse the values given by those respondents who do not have strong, well-grounded opinions (which is illustrated by wide ranges of fuzzy numbers). Respondents with such a manner of assessment represent a completely different perception of subjective values of verbal categories. Similarly, those respondents who have strong, well-grounded opinions, which are manifested in narrow ranges of fuzzy numbers, need a different approach in the interpretation of measurement results. They demonstrate a wider knowledge and are more focused on the differentiation between latent values behind verbal categories.

Acknowledgments: The study was partly conducted within the research project entitled 'Households' equipment with durable goods in statistical analysis and econometric modelling of material wellbeing.' Project no. 2018/29/B/HS4/01420 is financed by the National Science Centre, Poland.

References

Abdi H., (2010), Guttman Scaling, in: N. Salkind, (ed.), *Encyclopedia of Research Design*, Sage, Thousand Oaks.

Ackerman T., (2005), Multidimensional Item Response Theory Models, in: B. Everitt, D. Howell, (ed.), *Encyclopedia of Statistics in Behavioural Science, vol. 3*, Wiley, Chichester.

Arguelles Mendez L., (2016), From Fuzzy Sets to Linguistic Variables, in: L. Arguelles Mendez, (ed.), *A Practical Introduction to Fuzzy Logic using LISP*, Springer, Berlin, 169–228. DOI: 10.1007/978–3–319–23186–0_6.

Ark van der A., (2012), New Developments in Mokken Scale ... in R., *Journal of Statistical Software*, 48(5). DOI: 10.18637/jss.v048.i05.

Brzezińska J., (2016), Modele IRT i modele Rascha w badaniach testowych, in: K. Jajuga, M. Walesiak, (ed.), *Taksonomia 27: Klasyfikacja i analiza danych – teoria i zastosowania*, 49–57.

Combrinck C., (2018), *The use of Rasch Measurement Theory to Address Measurement and Analysis Challenges in Social Science Research,* PhD thesis, University of Pretoria, https://repository.up.ac.za/bitstream/handle/2263/67982/Combrinck_Use_2018.pdf?sequence=1&isAllowed=y (accessed 20.11.2019).

DeMars C., (2018), Classical Test Theory and Item Response Theory, chapter 2, in: P. Irwing, T. Booth, D. Hughes, (ed.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development,* Wiley, New York, 49–73. DOI: 10.1002/9781118489772.ch2.

DePaoli S., Tiemensma J., Felt, J., (2018), Assessment of health surveys. Fitting a multidimensional graded response model, *Psychology, Health & Medicine*, 23(1), 1299–1317. DOI: 10.1080/13548506.2018.1447136.

Diener E., Oishi S., Tay L., (2018), *Handbook of Well-Being*, DEF Publishers, Salt Lake City.

Edelen M., Reeve B., (2007), Applying Item Response Theory (IRT) Modelling to Questionnaire Development, Evaluation and Refinement, *Quality of Life Research*, 16(5), 5–18. DOI: 10.1007/s11136–007–9198–0.

European Commission, (2017), Qualitative Analysis. Verticals and Environments, chapter 5, in: *Identification and Quantification of Key Socioeconomic Data to Support Strategic Planning for the Introduction of 5G in Europe*, Publications Office of the European Union, Luxembourg, 6–7.

Fattore M., Maggino F., Greselin F., (2011), Socioeconomic Evaluation with Ordinal Variables. Integrating Counting and POSET Approaches, *Statistica and Applicazioni,* Special issue, 31–42.

FisPro: An Open Source Portable Software for Fuzzy Inference Systems, (2018), https://www.fispro.org/en/documentation (accessed 20.11.2019).

Guttman L., (1944), A Basis for Scaling Qualitative Data, *American Sociological Review*, 9(2), 139–150. DOI: 10.2307/2086306.

Guttman L., Stouffer S., Suchman E., Lazarsfeld P., Star S., Clausen J., (1950), *Measurement and Prediction*, Princeton University Press, Princeton.

Hambleton R., Swaminathan H., (1991a), Assumptions of Item Response Theory, in: R. Hambleton, H. Swaminathan, *Item Response Theory. Principles and Applications*, Springer, Berlin.

Hambleton R., Swaminathan H., (1991b), *Item Response Theory. Principles and Applications*, Springer, Berlin.

Hambleton R., Swaminathan H., (1991c), Shortcomings of Standard Testing Methods, in: R. Hambleton, H. Swaminathan, *Item Response Theory. Principles and Applications*, Springer, Berlin, DOI: 10.1007/978-94-017-1988-9.

Hartig J., Hoehler J., (2009), Multidimensional IRT Models for the Assessment of Competencies, *Studies in Educational Evaluation*, 35(2–3), 57–63. DOI: 10.1016/j.stueduc.2009.10.002.

Hofmann R., (1979), On Testing a Guttman Scale for Significance, *Educational and Psychological Measurement*, 39(2), 297–301. DOI: 10.1177/001316447903900206.

Immekus J., Snyder K., Ralston P., (2019), Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research, *Frontiers in Education*, 4, 1–15. DOI: 10.3389/feduc.2019.00045.

Jabrayilov R., Emons W., Sijtsma K., (2016), Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment, *Applied Psychological Measurement*, 40(8), 559–572. DOI: 10.1177/0146621616664046.

Jefmański B., (2014), Application of Rating Scale Model in Conversion of Rating Scales' Points to the Form of Triangular Fuzzy Numbers, *Folia Oeconomica Stetinensa*, 14(2), 7–18. DOI: 10.1515/foli-2015-0010.

Jumailiyah M., (2017), Item Response Theory: A Basic Concept, *Educational Research and Reviews*, 12(5), 258–266. DOI: 10.5897/ERR2017.3147.

Kacprzyk J., Roubens M., (ed.), (1988), *Non-Conventional Preference Relations in Decision Making*, Springer, Berlin.

Kappenburg -ten Holt J., (2014), *A Comparison Between Factor Analysis and Item Response Theory Modelling in Scale Analysis, PhD thesis, University of Groningen*, https://www.rug.nl/research/portal/files/13080475/20140623_Gmw_TenHolt.pdf (accessed 20.11.2019).

Kong, N., (2018), *Numerical Comparisons across General Total Score, Total Score, and Item Response Theory*, https://www.researchgate.net/publication/323931094_Numerical_Comparisons_across_General_Total_Score_Total_Score_and_Item_Response_Theory. DOI: 10.13140/RG.2.2.21519.89768.

Linacre J., (2000), Comparing and Choosing between Partial Credit Models (PCM) and Rating Scale Models (RSM), *Rasch Measurement Transactions*, 14(3), 768.

Linacre J., (2002), Optimizing Rating Scale Category Effectiveness, *Journal of Applied Measurement*, 3(1), 85–106.

Lynn P., (2019), Applying Prospect Theory to Participation in a CAPI/WEB Panel Survey, *Public Opinion Quarterly*, 83(3), 559–567. DOI: 10.1093/poq/nfz030.

Magno C., (2009), Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data, *The International Journal of Educational and Psychological Assessment*, 1(1), 1–11.

Maydeu-Olivares A., Cai L., Hernandez A., (2011), Comparing the Fit of Item Response Theory and Factor Analysis Models, *Structural Equation Modelling*, 18(3), 333–356. DOI: 10.1080/10705511.2011.581993.

Michalos A., (ed.) (2014), *Encyclopedia of Quality of Life and Wellbeing Research*, Dordrecht, Berlin.

Mokken R., (1971), *A Theory and Procedure of Scale Analysis with Applications in Political Research*, de Gruyter, Berlin.

NAP, (2017), *Improving Motor Carrier Safety Measurement*, The National Academies Press, Washington.

ONS, (2019a), *Measuring National Wellbeing in the UK: International Comparisons*, Office for National Statistics, https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/measuringnationalwellbeing/internationalcomparisons2019#personal–well–being (accessed 20.11.2019).

ONS, (2019b), Personal Wellbeing in the UK QMI, https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingintheukqmi#methodology–background (accessed 20.11.2019).

Prieler J., (2007), So Wrong for so Long. Changing our Approach to Change, *The Psychologist,* 20(12), 730–732.

Raykov T., Marcoulides G., (2016), On the relationship between classical test theory and item response theory. From one to the other and back, *Educational and Psychological Measurement,* 76(2), 325–338. DOI: 10.1177/0013164415576958.

Roubens M., Vincke P., (1988), Fuzzy Possibility Graphs and Their Application to Ranking Fuzzy Numbers, in: J. Kacprzyk, M. Roubens, (ed.), *Non-Conventional Preference Relations in Decision Making* , Springer, Berlin, 119–128.

Royal K., Ellis A., Ensslen A., Homan A., (2010), Rating Scale Optimization in Survey Research: An Application of the Rasch Rating Scale Model, *Journal of Applied Quantitative Methods*, 5(4), 607–617.

Saisana M., (2014), Composite Indicator(s), in: A. Michalos, (ed.), *Encyclopedia of Quality of Life and Wellbeing Research*, Dordrecht, Berlin, 1156–1161.

Schnorr-Bäcker S., (2018), The Possibilities and Limitations of Measuring Prosperity and Wellbeing in Official Statistics, in: *Essays by the Members of the Scientific Advisory Board Government Strategy on Wellbeing in Germany*, German Government, Berlin, 74–85.

Sijtsma K., Ark van der A., (2017), A Tutorial on how to do a Mokken Scale Analysis on your Test and Questionnaire Data, *British Journal of Mathematical and Statistical Psychology*, 70(1), 137–185. DOI: 10.1111/bmsp.12078.

Sireci S., Thissen D., Wainer H., (2005), On the Reliability of Testlet-Based Tests. *Journal of Educational Measurement*, 28(3), 237–247. DOI: 10.1111/j.1745–3984.1991.tb00356.x.

Tinsley H., Brown S., (2000), *Handbook of Applied Multivariate Statistics and Mathematical Modelling,* Academic, New York.

Tov W., Diener E., (2009), Culture and Subjective Wellbeing, in: S. Kitayama, D. Cohen, (ed.), *Handbook of Cultural Psychology*, Guilford, New York, 691–713. DOI: 10.1007/978-90-481-2352-0_2.

Wainer H., Kiely G., (1987), Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement*, 24(3), 185–201. DOI: 10.1111/j.1745-3984.1987.tb00274.x.

Walesiak M., Gatnar E., (ed.), (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.

Watson R., Egberink I., Kirke L., Tendeiro J., Doyle F., (2018), What are the Minimal Sample Size Requirements for Mokken Scaling? An Empirical Example with the Warwick Edinburgh Mental Wellbeing Scale, *Health Psychology and Behavioural Medicine*, 6(1), 203–213. DOI: 10.1080/21642850.2018.1505520.

Wind S., (2017), An Instructional Module on Mokken Scale Analysis, *Educational Measurement: Issues and Practice*, 36(2), 50–66. DOI: 10.1111/emip.12153.

Yau H., Yao W., (2011), *Optimizing Distribution of Rating Scale Category in Rasch Model, Paper for 76th Annual and the 17th International Meeting of the Psychometric Society*, The Hong Kong Institute of Education, Hong Kong, https://pdfs.semanticscholar.org/3b8a/a68885c1e3f2be7e771b4335 fcc47d3f2bd5.pdf (accessed 20.11.2019).

Zadeh L., (1975), The concept of a linguistic variable and its application to approximate reasoning, *Information Sciences*, 8(3), 199–249. DOI: 10.1016/0020–0255(75)90036–5.

Zalnezhad E., Sarhan A., (2014), Fuzzy Modelling to Predict the Adhesion Strength of TiN Ceramic Thin Film Coating on Aerospace, in: L. Ye, (ed.), *Recent Advances in Structural Integrity Analysis*, Woodhead Publishing, Sawston, Cambridge, 239–244.

Zamri N., Abdullah L., (2014), A New Positive and Negative Linguistic Variable of Interval Triangular Type-2 Fuzzy Sets for MCDM, in: T. Herawan, R. Ghazali, M. Deris, (ed.), *Recent Advances on Soft Computing and Data Mining*, Springer, Cham, 69–78.

Zhu W., (2002), A Confirmatory Study of Rasch-based Optimal Categorization of a Rating Scale, *Journal of Applied Measurement*, 3(1), 1–15.