Using Item Response Theory to Analyze Properties of the Leadership Practices Inventory

Van der Stoop
University of Amsterdam,
Faculty of Psychology

Abstract

Paper examines the psychometric properties of the Leadership Practices Inventory (LPI) (Kouzes & Posner, 1993) in a framework of Item Response Theory (IRT). LPI assesses five dimensions (i.e. leadership practices) of the "neocharismatic" (House & Aditya, 1997) (or visionary and transformational) leadership and consists of 30 items. IRT is a model based theory that relates characteristics of questionnaire items (item parameters) and characteristics of individuals (latent variables) to the probability of choosing each of the response categories. IRT item parameters are not dependant on the sample of respondents to whom the questions were administered. Moreover, it does not assume that the instrument is equally reliable for all levels of the latent variable examined. Samejima's (1969) Graded Response Model was used to estimate LPI item characteristics, such as the item difficulty and item discrimination power. The results show that some items are redundant, in a sense that they contribute little to the overall precision of the instrument. Moreover, LPI seems to be most precise and reliable for respondents with low to medium usage of leadership practices, while it becomes increasingly unreliable for high-quality leaders. These findings suggest that LPI can best be used for training and development purposes, but not for leader selection purposes.

Keywords:
Leadership, Leadership practices, Item Response Theory, Questionnaire, Reliability

Introduction

The quality of leadership research depends on the quality of measurement. If various aspects of leadership phenomena are not measured properly, wrong conclusions may be drawn about relationships between them. Measurement precision of an instrument is crucial for the success of the inferences and decisions based upon that instrument, whether purpose is academic theory building or practical leader development effort.

Accurate and reliable measurement of leadership phenomena forms a foundation of leadership research. It is vital for discerning relationships between leadership and other managerial, economical, social and psychological phenomena. For example, how does leadership affect levels of organizational commitment or productivity? It is also critical for examining and evaluating the effectiveness of various leadership traits, competences or styles. It is important for theory building, but even more so for theory testing, and modification.

 ISSN 2047-3745

Furthermore, in recent years various measures of leadership are being adopted by organizations and individuals for the purpose of leadership training and development. They provide conceptual framework, measure individual's performance on different leadership subdimensions and allow her to compare her results with that of a reference group (company, industry or national average). Leadership inventories and questionnaires are also used to measure the success of leadership development initiatives and effects of training interventions (individual's and group's progress is examined by comparing questionnaire scores before and after the intervention). Some organizations are even using them for leader selection, promotion and compensation (Hughes, Ginnett, & Curphy, 1999).

This leads to a concomitant increase in the need for researcher and practitioner to evaluate the quality and measurement precision of the instruments used to estimate various managerial and leadership skills and competences. Various statistical more or less sophisticated statistical techniques are available for this purpose. From Cronbach's alpha to structural equation modeling, they all provide some insights into the properties of the measurement instruments.

However, the purpose of this article is to evaluate accuracy and reliability of well known leadership instrument – Leadership Practices Inventory (Kouzes & Posner, 1987; 1993) – by using *item response theory (IRT)* (Lord & Novick, 1968; Van der Linden & Hambleton, 1997). Because IRT is fairly recent and not yet widely known technique in the field of leadership research, the secondary purpose is to explain the basics of IRT, and demonstrate some of it's techniques that can be usefully applied in the field of leadership research and more specifically questionnaire development and testing.

Item response theory (IRT) presents an excellent methodology for evaluation of leadership instruments in this regard, given than unlike classical test theory (CTT) it does not assume that tests are equally precise across the full range of possible test scores. That is, rather than providing a point estimate of the standard error of measurement (SEM) for a leadership scale as in CTT, IRT provides a test information function (TIF) and a test standard error (TSE) function to index the degree of measurement precisions across the full range of the latent construct (denoted θ). Using IRT, leadership instruments can be evaluated in terms of the amount of information and precision they provide at specific ranges of test scores that are of particular interest. For example, in training and development applications, instruments that are equally precise across whole range of θ are desirable. However, for selection and promotion purposes measurement precision at the upper end of the θ continuum would likely be of main interest, and even a relatively large lack of precision at the lower end of the scale might be excused. Because many standardized tests tend to provide their highest levels of measurement precision in the middle range of scores, with declines in precision being seen at the high and low ends of the scale (Trippe & Harvey, 2002), it is quite possible that a test might be deemed adequate for assessing individuals scoring in the middle range of the scale, but unacceptably precise at the high or low ends (which, depending on the direction of the test's scales, may represent precisely the most relevant ranges of scores for leader selection or promotion purposes).

First section of the article will introduce basic concepts of IRT and some IRT models. Next, some advantages of the IRT over the classical test theory will be discussed. Second section will introduce the Leadership

 ISSN 2047-3745

Practices Inventory (Kouzes et al., 1987, 1993) and sample of MBA students to whom the questionnaire was administered. Major IRT assumptions as well as some techniques and recommendations for accessing the model fit will be described in this section as well. Next section will provide results of the preliminary tests of the model fit. The analysis of item parameter estimates and item and test information curves will follow. Major findings and conclusions will be summarized in the concluding section.

Overview of IRT

IRT has a number of advantages over CTT methods to access leadership competencies or skills. CTT statistics such as item difficulty (proportion of "correct" responses in dichotomously scored items), and scale reliability are contingent on the sample of respondents to whom the questions were administered. IRT item parameters are *not dependent on the sample* used to generate the parameters, and are assumed to be invariant (within a linear transformation) across divergent group within a research population and across populations (Reeve, 2002). In addition, CTT yields only a single estimate of reliability and corresponding standard error of measurement, whereas IRT models measure scale precision across the underlying latent variable being measured by the instrument (Cooke & Michie, 1997; Reeve, 2002). A further disadvantage of CTT methods is that a participant's score is dependent on the set of questions used for analysis, whereas IRT-estimated person's trait or ability level is independent of the questions being used. Because the expected participant's scale score is computed from their responses to each item (that is characterized by a set of properties), the IRT estimated score is sensitive to differences among individual response patterns and is a better estimate of the individual's true level on the ability continuum than CTT's summed scale score (Santor & Ramsay, 1998).

IRT is a probabilistic model for expressing the association between an individual's response to an item and the underlying latent variable (often called "ability" or "trait) being measured by the instrument (Reeve, 2002). The underlying latent variable in leadership research may be any measurable construct, such as transformational or transactional leadership, authoritative or participative leadership style or communication, teamwork skills and visionary skills. The latent variable, expressed as theta (θ), is a continuous *unidimensional* construct that explains the covariance among item responses (Steinberg & Thissen, 1995). People at *higher levels of θ have a higher probability of responding correctly or endorsing* an item.

IRT models are used for two basic purposes: to obtain scaled estimates of θ as well as to calibrate items and examine their properties (Lord, 1980). This study will focus on the latter issue.

Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of a positive response. A variety of IRT models have been developed for dichotomous and polytomous data. In each case, the probability of answering correctly or endorsing a particular response category can be represented graphically by an item (option) response function (IRF/ORF). These functions represent the nonlinear regression of a response probability on a latent trait, such as conscientiousness or verbal ability (Hulin, Drasgow, & Parsons, 1983).

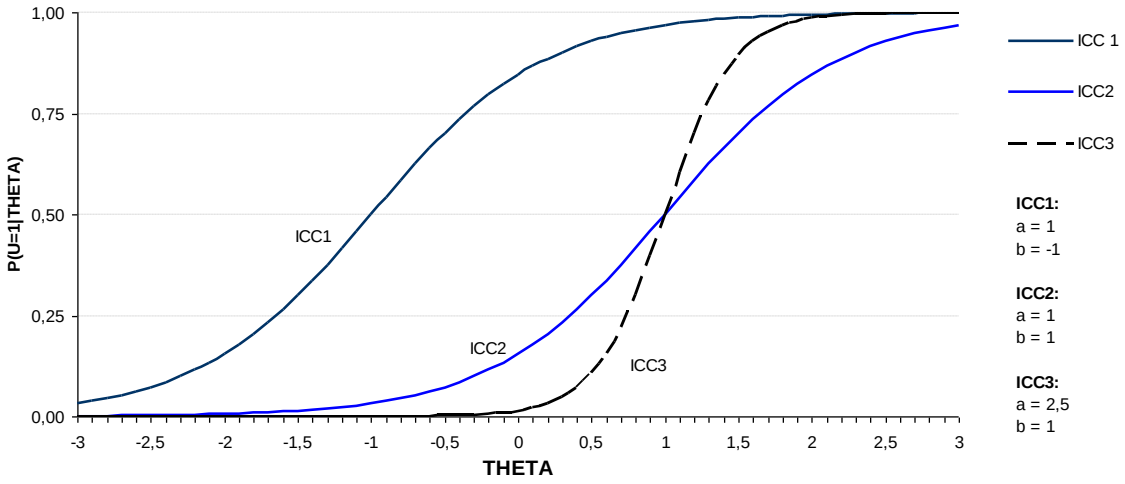
 ISSN 2047-3745

Each item is characterized by one or more model parameters. The item difficulty, or threshold, parameter *b* is the point on the latent scale θ where person has a 50% chance of responding positively or endorsing an item. Items with high thresholds are less often endorsed (Van der Linden et al., 1997). The slope, or discrimination, parameter *a* describes the strength of an item's discrimination between people with trait levels (θ) below and above the threshold *b.* The *a* parameter may also be interpreted as describing how strongly an item may be related to the trait measured by the scale. It is often thought of and is linearly related (under some conditions) to the variable loading in a factor analysis (Reeve, 2002).

To model the relation of the probability of a correct response to an item conditional on the latent variable θ, trace lines, estimated for the item parameters, are plotted. Most IRT models in research assume that the normal ogive or logistic function describes this relationship accurately and fits the data. The trace line (called the item characteristic curve, ICC or item response function IRR) can be viewed as the regression of item score on the underlying variable θ (which is usually assumed to have standardized normal distribution with mean 0 and standard deviation 1) (Lord, 1980).

Figure 1 presents item characteristic curves for three dichotomous items (scored 0 or 1). Items 1 and 2 have same *a* parameters but different thresholds (*b)* with item 2 being more difficult, that is, having lower probability of endorsement for each level of θ. Items 2 and 3 have equal thresholds, but differ in discrimination power (*a).* Item 3 is able to better discriminate between respondents than item 2.

Figure 1: Example of item characteristics curves (ICCs) for three items



The collection of item characteristic curves forms the scale; thus the sum of the probabilities of the correct response of the ICCs yields the test characteristic curve.

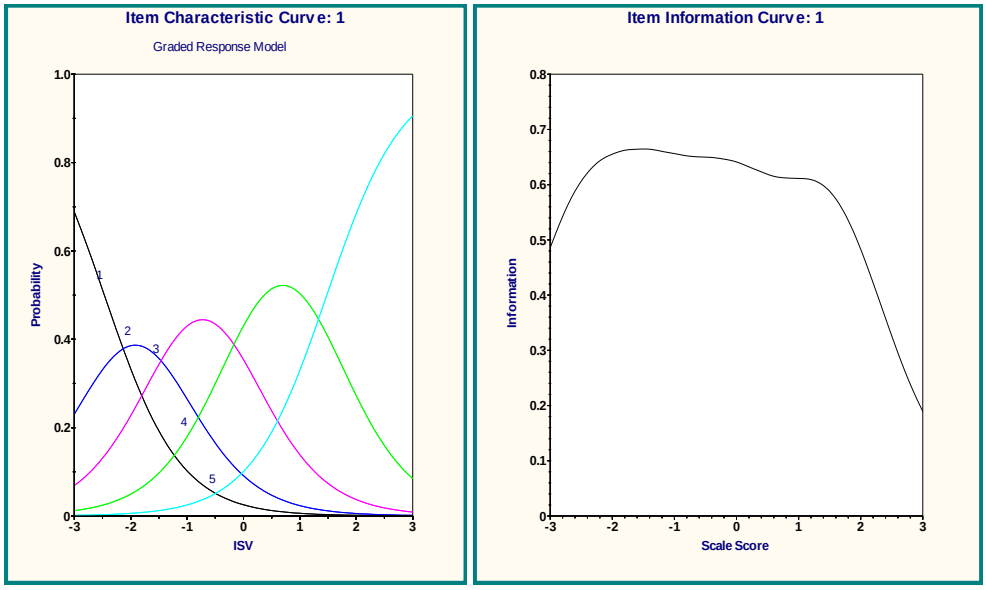
There exist numerous IRT models that differ in the type and number of item parameters estimated as well as in their suitability for different types of data. Rasch (1960) or one parameter model is suitable for

dichotomous data and estimates only items thresholds (*b)*. Two parameter model additionally estimates item difficulties (*a*). This model is represented on Figure 1. Three parameter model includes additional *guessing* parameter which estimates the lower asymptote of the ICC, and accounts for the fact that sometimes even people with low amounts of latent ability endorse an item.

For ordered polytomous data, that is questions with three or more response categories, Samejima's (1969; 1997) Graded Response Model (GRM) is most frequently used. A response may be graded on a range of ranked options, such as a five-point Likert-type scale used in this study. GRM is based on the logistic function giving the probability that an item response will be observed in category k or higher. Trace lines model the probability of observing each response alternative (k) as a function of the underlying construct (Steinberg et al., 1995). In other words, graded model estimates each response category's probability of endorsement at each possible value of θ.

Figure 2: Graded response model (1 item with 5 response categories).

a) b)



The slope a_i varies by item *i*, but within an item, all option response trace lines (ORFs) share the same slope (discrimination). This constraint of equal slope for responses within an item keeps trace lines from crossing, thus avoiding negative probabilities. The threshold parameters b_{ik} vary within the constraint $b_{k-1}<b_k<b_{k+1}$. The value of $b_{k-1}$is the point on the θ-axis at which the probability passes 50% that the response is in category k or higher (Thissen, 1991).

Figure 2a presents an example of one graded item with 5 response categories from this study. Respondent has to answer how frequently she "describes to others the kind of future she would like for the team to create

 ISSN 2047-3745

together." 1 denotes "*rarely, almost never"* and 5 means "*frequently, always"*. For example, at threshold b_3= *-0,08* there exist 50% probability that the respondent with just a little bit under average level of leadership ability will endorse categories 4 or 5. For respondent with θ= -1 there is approximately 10% probability that she will choose category 1, 24% probability that she will choose category 2, 44% probability for 3, 20% for 4 and 2% probability that she will choose category 5.

Samejima's parameter b_k can be resolved into an item location parameter b_j *(*average b_k*)* and category parameter c_k. Parameter b_j thus becomes indicator of average item difficulty that can be compared with other items in the scale, while c_k denotes relative position of each category in reference to b_j (Muraki & Bock, 2002).

For each item in the scale, as well as for the scale itself, *information function* can be calculated. It is an index indicating the range of ability level θ over which an item or scale is most useful for distinguishing among individuals. In other words, the information function characterizes the precision of measurement for persons at different levels of the underlying latent construct, with higher information denoting more precision. Example of *item information function* is presented in the Figure 2b. The shape of the item's information function is dependent on the item parameters. The higher the item's discrimination, the more peaked the information function will be; thus higher discrimination parameters provide more information about individuals whose trait levels (θ) lie near the item's threshold value. The item's difficulty parameter(s) determine the location of the item information function and it's peak (Reeve, 2002). For example, item from Figure 2b provides most information for respondents at θ = -2 and least information for respondents with θ > 2. With the assumption of the local independence the item information functions can be summed across all the items in the scale to form the test information curve (Lord, 1980).

At each level of the underlying construct θ, the information function is approximately equal to the expected value of the inverse of the squared standard error of the θ-estimates (Lord, 1980). The smaller the standard error of measurement (SEM), the more information or precision the scale provides about θ.

Advantages of IRT over classical test theory

Application of IRT in survey type of leadership research has several advantages over classical test theory. These are summarized in Table 1.

Table 1: Comparison of IRT and CTT

Classical Test Theory	Item Response Theory
Measures of precision fixed for all scores	Precision measures vary across scores
Longer scales increase reliability	Shorter, targeted scales can be equally reliable
Test properties are sample dependant	Test properties are sample free
Latent variable level estimated directly from the raw scores.	Latent variable level estimated from the raw scores, score patterns and item

 ISSN 2047-3745

	properties (thresholds and difficulties).
Comparing respondents requires parallel scales	Different scales can be placed on a common metric
Summed scores are on ordinal scale	Scores on interval scale

Unlike CTT item statistics, which depend fundamentally on the subset of items and persons examined, IRT item and person parameters are invariant. In CTT it is assumed that equal ratings on each item of the scale represent equal levels of the underlying trait (Cooke et al., 1997). Item response theory, on the other hand, estimates individual latent trait level scores based on all the information in a participant's response pattern. That is, IRT takes into consideration, which items were answered positively (high grade) and which negatively (low grade), and utilizes the difficulty and discrimination parameters of the items when estimating trait levels. Therefore, persons with the same summed score but different response patterns may have different IRT estimated latent scores (Reeve, 2002).

Consider, for example, two items from the leadership risk-taking scale "Challenging the process" (discussed in the next section): (1) "*I stay up-to-date on the most recent developments affecting our organization*," and (2) "*I experiment and take risks with new approaches to my work, even when there is a chance of failure*." It is obvious that second item is more "difficult" than the first one, in a sense that it implies more risk-taking behavior than first. Under CTT a person who scores 5 on first and 3 on second item will have the same score as a person who scores 3 on first and 5 on second item. Under most IRT models, the difficulty and discrimination power of those items would be taken into account, so the two respondents would have different score.

The correlation between summed score and IRT scores is usually quite high. However, summed scores are on a scale that assumes *the distance between any consecutive scores is equal*. It is a common practice to assume that, for example, the distance between category 1 and 2 on a 5 point Likert variable is *same* as the distance between 3 and 4 or between 4 and 5. IRT models are on an interval scale and distance between scores vary depending on the difficulty (and sometimes discriminating power) of the question (more exactly response categories in each question) (Reeve, 2002).

Furthermore, IRT allows the researches to "shape" the characteristics of a test by retaining items which provide desirable levels of information at specific points on the construct continuum. Next, IRT allows researchers to conduct rigorous tests of measurement equivalence across experimental groups. This is particularly important in cross-cultural research where groups are expected to show mean differences on the attribute being measured. IRT methods can distinguish item bias from true differences on the attribute measured, whereas CTT methods cannot (Kim, Cohen, & Park, 1995)).

Finally, IRT also facilitates computer adaptive testing. Items can be selected that provide the most information for each examinee. This can dramatically reduce time and costs associated with test administration (Hulin, Drasgow, & Parsons, 1983).

 ISSN 2047-3745

IRT and Leadership

Despite the advantages offered by IRT over the older CTT-based methods of assessing instrument performance, it is relatively modestly used in the leadership research field. Some of the possible reasons are listed in the concluding section of the paper.

One example of successful application of IRT was reported in Craig and Gustafson (1998). Graded response model (Samejima, 1969) was used to reduce the length of the newly developed *Perceived leader integrity scale.* First, MULTILOG (Thissen, 2003) computer program was used to estimate item parameters and information functions for 43 items, that were found to adequately measure leader integrity construct. Next, the least desirable item, identified on the basis of its information function, was removed from the scale. The least desirable item was the item that provided less information in the θ range from -1.0 to 1.0 than any other item.

This method was repeated iteratively. After each deletion, MULTILOG was used to estimate item parameters and information functions for the remaining items. The test information function was also examined at each step and reduction of its precision noted. 12 items were removed from the scale with almost no degradation of its test information function. Further attempts to reduce the number of items caused a significant decrease in the amount of information provided by the test. As a result of IRT analysis, final scale was reduced to 31 items with almost no loss in its precision.

Method

Leadership practices inventory

There exist plethora of leadership questionnaires developed both by consultants and academics that assess leadership competences and skills of respondents. They differ in conceptual framework used to develop them, in number and substance of dimensions they measure, in rigorousness of their development and in their psychometric properties. However, in recent years instruments based on "neocharismatic leadership theories" like transformational leadership theory (Bass, 1985; Burns, 1978) and visionary leadership theory (Bennis & Nanus, 1985; Kouzes et al., 1987; Sashkin, 1988) are becoming increasingly popular, both with scholars and practitioners alike.

This study examines Leadership Practices Inventory (LPI) developed by Kouzes and Posner (1987). The inventory is based upon case study analyses of more than 1100 managers and their "personal best experiences." These written cases were supplemented with in-depth interviews conducted with thirty-eight middle to senior level managers (Kouzes et al., 1993). This qualitative analysis revealed a pattern of underlying and critical leadership actions and behaviors. They were grouped into five behaviors (practices) that are common to successful leaders:

 ISSN 2047-3745

1. Challenging the Process (CP): searching for challenging opportunities, questioning status quo, experimenting and taking risks.
2. Inspiring a Shared Vision (ISV): envisioning an exciting future and enlisting others to pursue that future.
3. Enabling Others to Act (EOA): fostering collaboration, empowering and strengthening others.
4. Modeling the Way (MW): consistently practicing one's own espoused values, setting the example, planning small wins.
5. Encouraging the Hearth (EH): giving positive feedback, recognizing individual contributions and celebrating team accomplishments.

Most of the behaviors measured by LPI could be classified as transformational under transactional /transformational leadership paradigm (Bass, 1985, 1997). Kouzes and Posner's Leadership practices model (1987) is of a common genre and has several common characteristics with other "neocharismatic" (transformational, charismatic and visionary) leadership theories (House et al., 1997). First, they all "attempt to explain how leaders are able to lead organizations to attain outstanding accomplishments such as the founding and growing of successful entrepreneurial firms, corporate turnarounds in the face of overwhelming competition, military victories in the face of superior forces, and leadership of successful social reform for independence from colonial rule or political tyranny. Second, these theories also attempt to explain how certain leaders are able to achieve extraordinary levels of follower motivation, admiration, respect, trust, commitment, dedication, loyalty, and performance. Third, they stress symbolic and emotionally appealing leader behaviors, such as visionary, frame alignment, empowering, role modeling, image building, exceptional, risk taking, and supportive behaviors, as well as cognitively oriented behavior, such as adapting, showing versatility and environmental sensitivity, and intellectual stimulation. Finally, the leader effects specified in these theories include follower selfesteem, motive arousal and emotions, and identification with the leader's vision, values, and the collective, as well as the traditional dependent variables of earlier leadership theories: follower satisfaction and performance" (House, Aditya, 1997).

Leadership Practices Inventory (LPI) measures each of the five dimensions of (transformational) leadership with 6 statements. Individuals have to respond how frequently they employ specified behavior. Example statements are: *I seek out challenging opportunities that test my skills and abilities* (CP); *I am contagiously excited and enthusiastic about future possibilities* (ISV); *I involve others in planning the actions we will take* (EOA*); I am consistent in practicing the values I espouse* (MW); *I praise people for a job well done* (EH). Each statement is cast on a five-point Likert scale, with higher value representing greater use of the measured leadership behavior. Therefore each leadership practice can be scored in the range from 6 to 30 points. Extensive testing revealed that instrument exhibited sound psychometric properties (Kouzes, Posner, 1993; 1990). LPI has become quite popular among practitioners and has been widely used by business organizations in various parts of the world, primarily for management development purposes.

Sample

Where random sampling is problematic (as it is in management research), one way to increase the generalizability of findings is to deliberately sample for heterogeneity (Mark & Cook, 1984). By

 ISSN 2047-3745

intentionally selecting subjects who come from diverse geographic and cultural settings, the researcher can determine whether selected theory or a model accurately describes the actions of individuals across these divergent contexts. In order to achieve this, LPI was administered to MBA students in six countries from five continents: USA, India, Nigeria, South Korea, Argentina and Slovenia. English version of the questionnaire was used in first three countries while for the latter three countries LPI was translated to indigenous language. Sample sizes ranged from 110 in USA to 162 in Argentina. Percentage of females ranged form 31% in Nigeria to 48% in USA. After discarding 2 respondents that exhibited 0 variance (eg. by selecting "5" for all item responses), samples were pooled in one sample consisting of 801 respondents. Females represented 39% of sample. Half of the respondents had less than 10 years of work experience. Reliability of the instrument scales (Cronbach's alpa) was modest, ranging from .62 to .72.

Respondents belonged to different cultures with different "worldviews," customs, religions and different levels of economic development. By including them into one sample we were able to "randomize" background variables and assure higher validity of our conclusions than if the sample was drawn from just one culture (usually USA).

IRT assumptions

Before applying IRT model and estimating model fit and item parameters it is necessary to check if the assumptions on which item response theory is based hold in the actual sample. Most IRT models assume that a response to any one item is unrelated to other item reposes if the latent trait is controlled for (*assumption of local independence)* (Lord et al., 1968). Consequently, IRT models assume that the latent trait construct space is either strictly unidimensional, or as a practical mater, dominated by a general underlying factor (*assumption of unidimensionality)* (Trippe et al., 2002). For test using many items, this assumption might be unrealistic; however Cooke and Michie (1997) report that IRT models are moderately robust to departures from unidimensionality. More specifically, Kirisci, Hsu and Yu (2001) investigated the robustness of *polynomial* item parameters using MULTILOG to violations of the unidimensionality assumption, concluding that (a) when data are multidimensional, a test length of more than 20 items and a sample size of over 250 are necessary to recover stable parameter estimates, and (b) when there is one dominant dimension with several minor dimensions, a unidimensional IRT model is likely justified.

The assumption of unidimensionality can be examined by comparing the ratio of the first to the second eigenvalue on exploratory factor analysis. Alternatively one can examine the amount of variance explained by the first factor. Reckase (1979) recommended that the first factor account for at least twenty percent of the variance in order to obtain stable item parameters.

Model fit and item parameter estimations

MULTILOG (Thissen, 2003) was used to estimate Samejima's (1969) Graded Response Model parameters for items in each of the five scales separately. As was mentioned earlier, GRM is one of the most suitable and

 ISSN 2047-3745

well known models for examining Likert-type variables, used in LPI questionnaire. MULTILOG is one of the most used programs for estimation of polytomous IRT models.

The fit of the parameters obtained from each scale was evaluated using the graphical and statistical procedures. Fit plots are one of the most widely used graphical methods for examining model-data fit. Ideally, one would compare item/option response functions, estimated from a calibration sample, to empirical proportions of positive responses obtained from a cross-validation sample. However, in many applications, a cross-validation sample is not used (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Statistical tests of goodness of fit (i.e., χ^2 fit statistics) are probably the most widely used in applied research. Unfortunately, they are often viewed as inconclusive evidence of adequate fit because of their sensitivity to sample size and their insensitivity to certain forms of misfit. To avoid these problems, Drasgow et al(1995) recommend that the χ^2 statistic should be computed for pairs and triples of items. Pairs and triples of items with similar misfits will have large χ^2 statistics. To facilitate comparisons of χ^2 based on different sample sizes, Drasgow et al. (1995) advocated reporting χ^2, adjusted for sample size (say, 3000) and divided by their degrees of freedom. Based on numerous studies, they found that good model-data fit is associated with adjusted χ^2 to degrees of freedom *ratios of less than 3* for item singles, doubles and triples. Large ratio statistics for doubles and triples may indicate violations of local independence or unidimensionality. MODFIT (Stark, 2002) computer program (Excel macro) was used to obtain graphical plots and χ^2 to degrees of freedom ratios.

Results and discussion

Unidimensionality and local independence

To test the assumption of unidimensionality exploratory factor analysis was performed for each of the 5 leadership scales. Results are presented in Table 1. Percentage of variance explained by the first factor exceeds Reckase's (1979) criterion of 20% for all five scales. However, clear and dominant first factor is present only for first four scales. Two items of the Encouraging the Hearth scale (l05 and l25) load highly on the separate second factor that could be labeled *Celebrating accomplishments.*

Table 2: Eigenvalues and percentage of variance explained by first factors for 5 leadership practices scales

	Eigenvalue of the first factor	% of variance explained by the first factor	Eigenvalue of the second factor (EV2)	Ratio: EV1/EV2
Challenging the Process	2,39	39,79	0,90	2,64
Inspiring the Shared Vision	2,53	42,10	0,80	3,14
Enabling Others to Act	2,15	35,74	0,96	2,23
Modeling the Way	2,26	37,59	1,00	2,27
Encouraging the Hearth	2,40	39,98	1,20	1,99

 ISSN 2047-3745

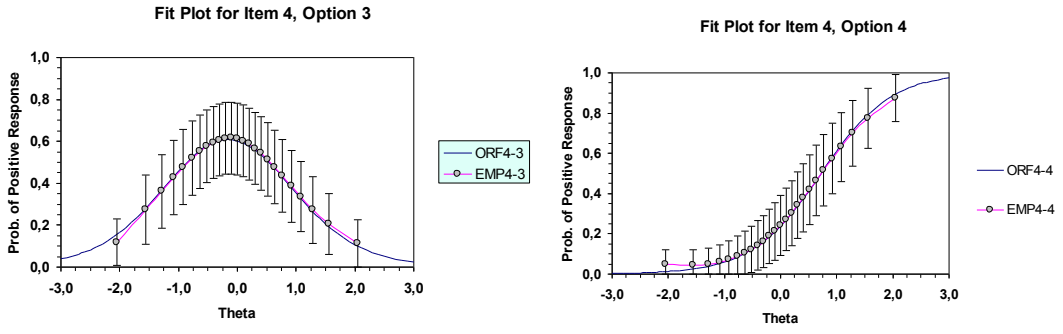
These results were confirmed with confirmatory factor analysis (CFA) using Lisrel 8.5 (Jöreskog & Sörbom, 1993) computer program. CFA allows for testing of hypothesized factor structure simultaneously. All parameters (factor loadings) were significant and in the right direction. Goodness of fit indices shoved reasonable, although not very exact fit (χ^2 = 1435; df=394; CFI=.94; NNFI=.93; GFI=.89; RMSEA =.058). However, when error covariance of the two items of the Encouraging the Hearth scale (l05 & l25) was set free, the fit increased significantly (χ^2 = 1214; df=394; CFI=.95; NNFI=.95; GFI=.91; RMSEA =.051). Therefore assumption of local independence does not hold for item two items of EH scale.

To sum, results show that four scales satisfy the assumptions of IRT. Fifth scale (EH) violates the assumption of unidimensionality and local independence. However, past research [cit] shows that moderate violations of strict unidimensionality identified in EH scale are unlikely to exert an appreciable distorting effect on IRT parameter estimation. However, results obtained for EH scale should be interpreted with caution, and cross validation with other samples should be performed to confirm their validity.

Model fit

Graphical analysis of fit plots showed excellent fit between the category response functions estimated with graded response model and empirical proportions of endorsed responses for each category. Because of the prohibitively large number of plots (5 scales * 6 items * 5 response categories = 150 plots) only two sample plots are presented here (Figure 4). The solid thin line, labeled ORF, is a theoretical category (option) response function computed from a calibration sample. The line with gray dots, referred to as EMP, is an empirical item response function computed from same sample (ideally one should use cross-validation sample, but because of the relatively small sample size, splitting it was not deemed feasible). The vertical lines in each figure describe the approximate 95% confidence intervals for the empirical points. It can be seen that there is a close correspondence between the ORF and EMP curves, which suggests that the GRM model fits the data well. Similar fit was obtained for other items, except for some items of the Encouraging the Hearth scale, where fit was not so close, but still within the bounds of confidence intervals.

Figure 3: Example of fit plots for two response categories (4 an 5) for item l21



 ISSN 2047-3745

To further assess the model fit, frequency distribution and average χ^2 / *df* ratios for all five LPI scales are reported in Table 2. Following recommendation by Drasgow et al (1995) both unadjusted and adjusted (to the sample size of 3000) ratios are shown. The mean χ^2 to *df* ratio for first four scales is well below Drasgow et al (1995) recommended cutoff of 3 for both unadjusted and adjusted results. However, adjusted ratios pairs and triples of items for Encouraging the Heart scale exceed the suggested cutoff value (the values of ratios are 3,42 and 4,56 respectively). For this scale 4 out of 15 pairs of items, as well as 9 out of 20 triplets have adjusted χ^2 / *df* ratios higher than 3. This is in line with previous findings, which suggest that EH scale is not unidimensional. Although unadjusted χ^2 to *df* ratios do not exceed 2, overall results imply that EH scale may not produce the most accurate results when subject to IRT analysis. These results should thus be treated with caution.

Table 3: χ^2 fit statistics for the five scales of LPI

		χ^2/df ratios								**Adjusted (n=3000) χ^2/df ratios**							
Scale		<1	1<2	2<3	3<4	4<5	>5	M.	SD	<1	1<2	2<3	3<4	4<5	>5	M.	SD
	Singlets	6	0	0	0	0	0	*0,03*	0,03	6	0	0	0	0	0	*0,00*	0,00
CP	Doublets	8	7	0	0	0	0	*0,90*	0,40	8	5	1	1	0	0	*0,96*	1,17
	Triplets	14	6	0	0	0	0	*0,96*	0,22	14	4	2	0	0	0	*0,90*	0,76
	Singlets	6	0	0	0	0	0	*0,03*	0,02	6	0	0	0	0	0	*0,00*	0,00
ISV	Doublets	9	6	0	0	0	0	*0,91*	0,37	9	3	2	1	0	0	*0,87*	1,14
	Triplets	6	14	0	0	0	0	*1,10*	0,25	6	10	3	1	0	0	*1,36*	0,92
	Singlets	6	0	0	0	0	0	*0,01*	0,00	6	0	0	0	0	0	*0,00*	0,00
EOA	Doublets	9	6	0	0	0	0	*0,89*	0,36	9	3	3	0	0	0	*0,88*	0,94
	Triplets	8	12	0	0	0	0	*1,11*	0,31	8	9	1	0	2	0	*1,42*	1,16
	Singlets	6	0	0	0	0	0	*0,01*	0,01	6	0	0	0	0	0	*0,00*	0,00
MW	Doublets	8	6	1	0	0	0	*1,03*	0,41	8	3	3	0	1	0	*1,24*	1,39
	Triplets	7	13	0	0	0	0	*1,16*	0,32	7	7	3	3	0	0	*1,62*	1,18
	Singlets	6	0	0	0	0	0	*0,03*	0,02	6	0	0	0	0	0	*0,00*	0,00
EH	Doublets	6	7	1	0	0	1	*1,60*	1,80	6	3	2	1	1	2	*3,42*	6,66
	Triplets	2	14	0	1	3	0	*1,95*	1,23	2	6	3	3	2	4	*4,56*	4,62

Item parameter estimates

For each of the six items in the every leadership practice scale MULTILOG calculates discrimination parameter (*a*), four threshold parameters (b_{1-4}) and item information curve. It also plots category trace curves and item information functions (see Figure 2 as an example). Four b_{1-4} parameters were combined

 ISSN 2047-3745

into relative location parameter b_j. Table 3 presents these parameters for each item of the five LPI scales. In each scale, items are ordered by their relative location on θ continuum (that is by their relative "difficulty").

Table 4: Item slope coefficients (a) and relative locations (b_j) on leadership practices dimensions

Scale		**Items**					
	Item	**L21**	**L16**	**L06**	**L01**	**L11**	**L26**
CP	a	0,75	1,60	1,08	1,46	1,15	1,27
	bj	-1,83	-1,78	-1,70	-1,37	-1,13	-0,70
	Item	**L12**	**L22**	**L27**	**L17**	**L02**	**L07**
ISV	a	1,06	1,20	1,00	1,58	1,48	1,42
	bj	-1,63	-1,57	-1,46	-0,76	-0,61	-0,19
	Item	**L08**	**L13**	**L18**	**L03**	**L23**	**L28**
EOA	a	0,89	0,79	1,55	0,64	1,80	1,43
	bj	-3,94	-2,74	-2,36	-2,33	-1,88	-1,60
	Item	**L24**	**L09**	**L29**	**L19**	**L04**	**L14**
MW	a	1,18	0,97	1,24	1,13	1,35	1,13
	bj	-2,27	-2,24	-1,83	-1,57	-1,43	-1,22
	Item	**L15**	**L10**	**L20**	**L25**	**L05**	**L30**
EH	a	1,70	1,41	1,81	1,00	0,90	1,09
	bj	-2,19	-1,98	-1,76	-1,13	-1,07	-1,85

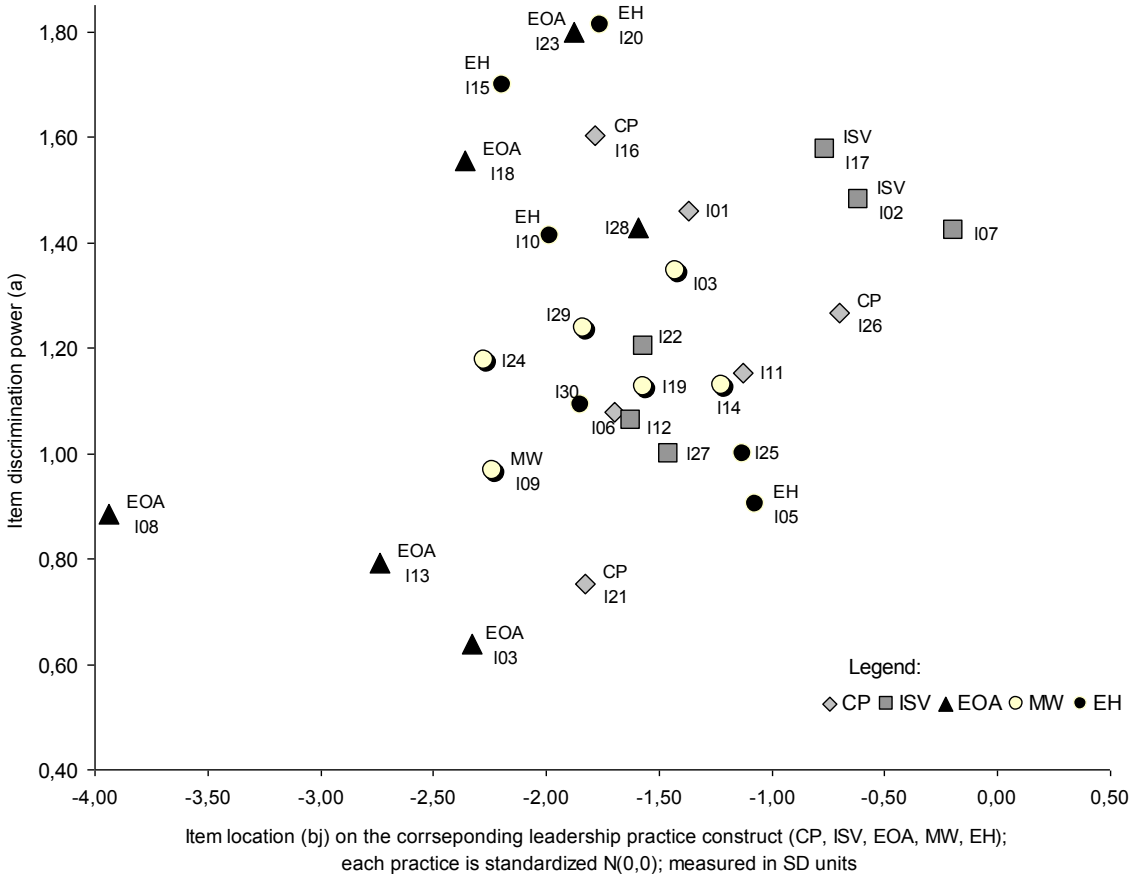
Note: Items are numbered as they appear in the questionnaire. For each scale items are ordered by their b_j (relative location) parameter.

It is evident that all items are located in the negative range of θ continuum. That means that they are able to accurately discriminate between respondents with relatively low levels of leadership ability, but not between respondents with high levels of leadership ability. The "easiest" item is l08: "I treat others with dignity and respect" from the Enabling Others to Act scale. Only really incompetent or tyrant leaders (total opposites of transformational leaders) will score low on this item. Therefore, this item is not very useful for discriminating between different degrees of transformational leadership. For most applications of the questionnaire it could easily be omitted from scale with almost no loss of instrument measurement precision.

Item parameters from Table 3 could be graphically mapped. Figure 5 presents the relationship between item location and discrimination parameters for all five LPI scales. It should be noted that although all scales are standardized and hence have same means and standard deviations, meaningful comparisons can only be made between items within the same scale, not across scales.

Figure 4: Map of location and item discrimination parameters for items of 5 LPI scales

 ISSN 2047-3745



Closer inspection of the Figure 5 reveals several items within each scale that have similar location on θ continuum, but substantially differ in their discriminating power. Example of such items are l21, l06 and l16 from Challenging the Process scale, l03 and l18 from Enabling Others to Act scale, l09 and l24 from Modeling the Way scale. Other items lie close together, being similar on both parameters, like l27, l12 and l22 from ISV scale. In both cases, removing some items with lower discrimination power will hardly affect instrument reliability and precision. They contribute little to measurement precision and do not discriminate among respondents.

IRT models test characteristics as a function of respondents' standing on θ, so the traditional notion of reliability is not meaningful in this context. That is, no single number can accurately describe the test's characteristics at all levels of θ. However, the "marginal" reliability index attempts to estimate a test's average reliability across the θ continuum (Thissen, 1986). This index will be used to examine the effect of removal of some of the items for the Challenging the Process scale (Table 4).

Table 5: Change in marginal reliability of CP scale with removal of some of the items

 ISSN 2047-3745

Scale description	MULTILOG marginal reliability estimate	% change in MR (compared to full scale)
Original CP scale with 6 items	0,717	-
Item l21 removed (5 item scale)	0,703	1,97%
Item l16 removed (5 item scale)	0,667	7,00%
Items l21 and l06 removed (4 item scale)	0,677	5,59%
Items l21, l06 and l26 removed (3 item scale)	0,601	16,21%

Marginal reliability (MR) of the CP scale equals to .717. From Figure 5 it can be seen that three items (l21, l06, l16) have similar b_j parameters in the range from -1.70 to -1.83. Because item l21 has lowest *a* (discrimination) it can be removed from the scale. MR drops slightly for less than 2%, from .717 to .703. However, if item l16 (that has high *a*) is removed, MR decreases for 7%, to .667. This decrease is higher than if two inferior items (l21 and l06) are jointly removed from the scale. Further removal of items results in sharp drop of MR, indicating that the optimal CP scale would consist of four items (l01, l11, l16, l26). Similar procedure could be employed for other scales of LPI.

Note that two items from EH scale that were found to violate the assumption of local independence have similar parameters. Removing one of them (e.g. l05) will solve the problem, while not significantly affecting scale properties.

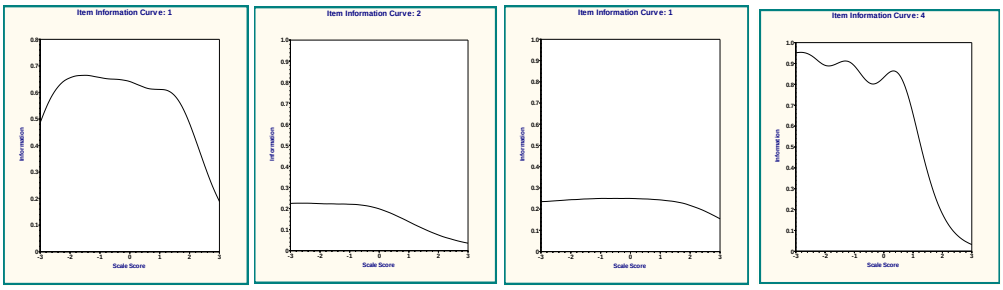
Item and test information curves

Depending on the combination of its parameters (a, b_{1-4}) each item has it's own *item information function (IIF).* Examples of some IIFs are given in Figure 6. It is clear that different items provide different amounts of information at different levels of θ. IIF for the first item in the Figure 6 (l02 from ISV scale) is the most common, typical IIF. It provides most information in the moderately low range of θ. Items two and three (l08, l21) provide relatively little information. Item 3 is equally imprecise over whole range of θ, while precision of item 3 declines at higher levels of θ. Item 4 (l14) provides large amounts of information (is highly precise) at lower ranges of θ, while its precision falls dramatically for higher θ levels.

Figure 5: Some examples of item information curves from leadership practices scales

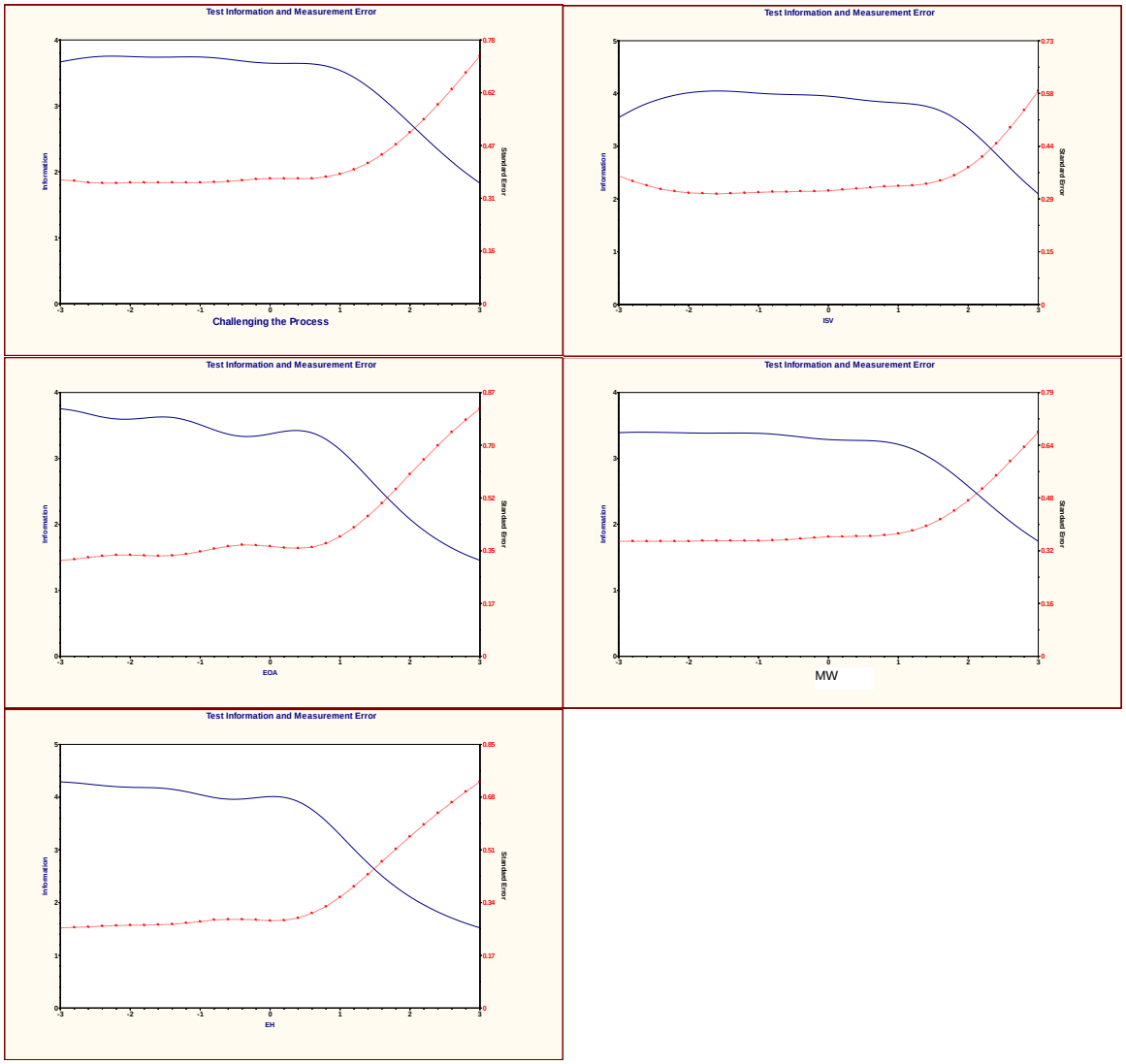
Item 1 Item 2 Item 3 Item 4

 ISSN 2047-3745



Item information functions for each of the five LPI scales were aggregated to form the test information functions (TIF) reported in Figure 7. Standard errors of measurement (SEM) for each scale are plotted on the same graphs (dotted lines). SEM is inversely related to test information function (TIF), hence as information for a given level of θ decreases the SEM increases. For each scale, θ denotes standardized level of leadership ability (practice) being measured by the scale (risk-taking, visionary, participative, integrity and exemplary leadership, recognition and praise, respectively).

Figure 6: Test information and standard error curves for five leadership practice scales



Note: The graphs represent the test information functions for the following leadership practices (from left to right and from top to bottom): Challenging the Process (CP), Inspiring the Shared Vision (ISV), Enabling Others to Act (EOA), Modeling the Way (MW), and Encouraging the Hearth (EH).

Test information functions for all five scales are relatively flat in the low range of θ continuum. They do not have distinct peaks, but generally provide most information for θ levels between -2 and 0 (measured in standard deviations). Their accuracy sharply decreases as θ increases over 1. Test information function for ISV scale appears to be most evenly spread out across whole range of θ continuum. In other words, this scale measures respondents with different levels of visionary leadership ability, from low to moderately high, with almost equal precision. However, precision of the scale sharply decreases and standard error of measurement

sharply increases when respondents with high visionary leadership ability are measured. Standard error of measurement consequentially increases from .30 in the low θ range to .60 for the θ>3.

EOA and EH scales exhibit the most prominent loss of information for θ levels higher than 0,5. EOA is at its most precise at the low levels of θ.

Conclusion

The purpose of this study was to assess and evaluate precision and accuracy of Leadership Practices Inventory across the whole range of leadership ability, in order to determine how useful the instrument is for research, training and development, self-evaluation and self-improvement, leader selection or promotion and compensation purposes. A modern measurement technique, Item Response Theory, was used to achieve this end. Article includes extensive section on IRT fundamentals, its properties, models and possible applications. However, it is by no means complete or adequate overview of IRT. Interested reader should consult Van der Linden & Hambleton (1997), Hambleton & Swaminathan (1985) or Lord (1980) for more thorough discussion of the topic.

Despite the advantages offered by IRT over the older CTT-based methods of assessing instrument performance it is relatively modestly used in the leadership research field. There are several reasons for this. First, IRT was developed within the framework of educational testing and so most of the literature and terminology is oriented towards that discipline (Hambleton et al., 1985). Second, major limitation of IRT is the complexity of the mathematical IRT models. Most researchers have been trained in classical test theory and are comfortable with reporting statistics such as summed scale scores, proportions correct, and Cronbach's alpha. Third, beyond the mathematical formulas, there are the complexities of the numerous IRT models themselves as to what circumstances are appropriate for IRT use and which model to choose. Fourth, there is not even a consensus among researchers as to the definition of measurement and which IRT models fit that definition. Finally, adding to the burden of confusion, the numerous available IRT software in the market are not user-friendly and often yield different results (Reeve, 2002).

Despite these limitations, practical applications of IRT in the field of leadership cannot be ignored. Only accurate instruments that correctly measure what they claim to measure can enhance out understanding of leadership phenomena and help us in the quest to select and develop better leaders. IRT dos not make existing psychometric techniques obsolete – it rather enhances them. Together, they allow us to obtain deeper insight into the nature and properties of a measurement instrument and offer us variety of tools to improve it.

Properties of the LPI were estimated using Samejima's (1969) graded response model. Results show that model fits well and is suitable for the analysis of the questionnaire. Examination of the parameters and test information functions revealed that all LPI scales perform well for respondents with low to modest levels of leadership ability. Test information functions in this θ range were relatively high and flat. However, scales become increasing unreliable in assessing the respondents with high levels of leadership ability.

 ISSN 2047-3745

Implications of these findings are several. LPI does not seem to be appropriate instrument for selection or promotion of high-quality leaders. In this case one focuses on the higher level of leadership ability scale (θ), where LPI is not able to reliability discriminate between good and excellent leaders. However, LPI is suitable for "screening" out bad or inferior leaders. Because its accuracy is stable over vide range of leadership ability (θ) it can be used for leadership development purposes, especially for lower and middle management, where one could not expect unproportionally high number of excellent leaders. In this case, it can reasonably well identify leadership strengths and weaknesses of the person, compare her score with reference groups scores and measure her progress in leadership ability (as a result of leadership development intervention or on-the-job learning).

Managers in modern companies are being flooded with various requests for participation in surveys, interviews or focus groups. They are becoming increasingly reluctant to answer all kinds of surveys, that they perceive are just stealing their precious time. That is why one of the objectives of instrument developers is to develop as short scales as possible for the given level of accuracy. In this way they reduce the burden on respondents, sustain their motivation, increase response rate and accuracy of the responses.

Study has demonstrated that LPI could be shortened without noticeable drop in its measurement precision or increase of the standard error of measurement. Items with low discrimination power, especially those that have similar thresholds as other items, do not contribute much to test information function and could therefore be safely removed from the scale. For example, removal of one such item for the Challenging the Process scale (l21, l06) reduced marginal reliability of scale from 0,717 to 0,703. Reduction of scale length for 20% caused decrease in marginal reliability of only 2%. This technique becomes even more powerful with large number of items in the scale. On the other hand, it could also be applied in the opposite direction. New items, located (b_j) in the higher end of leadership ability continuum (θ), could be added to questionnaire. This would improve accuracy of the scales, and hence improve their marginal reliability. Furthermore, because of the nice properties of IRT models (independence of person scores from item properties) person scores obtained with new, improved scale, would still be comparable to scores obtained with previous versions of the scale (only standard error of estimates would be lower).

Overall, IRT analysis shows that Leadership Practices Inventory appears to be moderately reliable instrument, better suited for leadership development than for leader identification, selection or promotion purposes. Its scales could be improved by removing some of the redundant items with low discrimination power (e.g. l21, l06, l09, l05, l03), and by adding new, more "difficult" items located in the upper part of the leadership ability (θ) continuum.

 ISSN 2047-3745

References

Bass, B. M. 1985. *Leadership and Performance Beyond Expectations*. New York: Free Press.

Bass, B. M. 1997. Does the Transactional - Transformational Leadership Paradigm Transcend Organizational and National Boundaries? *American Psychologist*, 52(2): 130-139.

Bennis, W., & Nanus, B. 1985. *Leaders: The Strategies for Taking Charge*. New York: Harper & Row.

Burns, J. M. 1978. *Leadership*. New York: Harper & Row.

Cooke, D. J., & Michie, C. 1997. An Item Response Theory Analysis of the Hare Psychopathy Checklist - Revised. *Psychological Assessment*, 9: 3-14.

Craig, S. B., & Gustafson, S. B. 1998. Percieved Leader Integrity Scale: An Instrument for Accessing Employee Perceptions of Leader Integrity. *Leadership Quarterly*, 9(2): 127-146.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. 1995. Fitting polytomous item response theory models to multiple choice tests. *Applied Psychological Measurement*, 19: 143-165.

Hambleton, R. K., & Swaminathan, H. 1985. *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

House, R. J., & Aditya, R. N. 1997. The Social Scientific Study of Leadership: Quo Vadis? *Journal of Management*, 23(3): 409-474.

Hughes, R. L., Ginnett, R. C., & Curphy, G. J. 1999. *Leadership : enhancing the lessons of experience* (3rd ed.). Boston, Mass.: Irwin/McGraw Hill.

Hulin, C. L., Drasgow, F., & Parsons, C. K. 1983. *Item Response Theory: Applications to Psychological Measurement*. Homewood, IL: Dow Jones Irwin.

Jöreskog, K. G., & Sörbom, D. 1993. *Lisrel 8: Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, S. H., Cohen, A. S., & Park, T. H. 1995. Detection of Differential Item Functioning in Multiple Groups. *Journal of Educational Measurement*, 32: 261 - 276.

Kirisci, L., Hsu, T., & Yu, L. 2001. Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*(25): 146 - 162.

 ISSN 2047-3745

Kouzes, J. M., & Posner, B. Z. 1987. *The Leadership Challenge: How to Get Extraordinary Things Done in Organizations*. San Francisco: Jossey - Bass.

Kouzes, J. M., & Posner, B. Z. 1993. *Psychometric Properties of the Leadership Practices Inventory*. San Diego: Pfeifer & Company.

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

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

Mark, M., & Cook, T. 1984. Design of randomized experiments and quasi-experiments. In L. Rutman (Ed.), *Evaluation Research Methods: A Basic Guide*, 2nd ed. Newbury Park, CA: Sage.

Muraki, E., & Bock, D. R. 2002. Parascale 4 Help File: SSI.

Reckase, M. D. 1979. Unifactor latent trait models applied to multifactor test: Results and implications. *Journal of Educational Statistics*(4): 207-230.

Reeve, B. B. 2002. An Introduction to Modern Measurement Theory.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 34(17).

Samejima, F. 1997. Graded Response Model. In W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. NY: Springer.

Santor, D. A., & Ramsay, J. O. 1998. Progress in the Technology of Measurement: Applications of Item Response Models. *Psychological Assessment*, 10: 345-359.

Sashkin, M. 1988. The Visionary Leader. In J. A. Conger, & R. A. Kanungo (Eds.), *Charismatic Leadership: The elusive factor in organizational effectiveness*: 122-160. San Francisco: Jossey-Bass.

Stark, S. 2002. MODFIT Computer program.

Steinberg, L., & Thissen, D. 1995. Item response theory in personality research. In P. E. Shrout, & S. T. Fiske (Eds.), *Personality research, methods, and theory: a festschrift honoring Donald W. Fiske*. Hilldale, NJ: Erlbaum.

Thissen, D. 1986. *MULTILOG: Item analysis and scoring with multiple category response models (Version 6)*. Mooresville, IN: Scientific Software.

Thissen, D. 1991. *Multilog User's Guide - Version 6*. Chicago, IL: SSI.

Thissen, D. 2003. Multilog, 7.03 ed. Chicago, IL: SSI.

Trippe, M. D., & Harvey, R. J. 2002. Item Response Theory Analysis of the IPIP Big-Five Scales.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). 1997. *Handbook of Modern Item Response Theory*. NY: Springer.