International Journal of Educational Research Open 2 (2021) 100054

Contents lists available at ScienceDirect

International Journal of Educational Research Open

journal homepage: www.elsevier.com/locate/ijedro

Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIW) adult playfulness scale ☆

Darrel R. Davis*, William Boone

Department of Educational Psychology, Miami University, Oxford, OH 45056, United States

ARTICLE INFO

Keywords:
OLIW evaluation
Playfulness instrument
Rasch measurement
Adult playfulness

ABSTRACT

It is important for instruments measuring playfulness constructs to be valid and reliable because these instruments help inform playfulness-related theories. This study uses Rasch psychometric techniques to evaluate the functioning of the OLIW instrument (Proyer, 2017a), a questionnaire that assesses four adult playfulness traits (Other-directed, Lighthearted, Intellectual, and Whimsical). Results confirm that the instrument measures four discrete traits. Analyses of the rating scale suggest acceptable scale functioning. The item hierarchy observed on each Rasch Wright Map provides new details regarding the definition of each construct (e.g. what it means for a person to be at various locations along each construct). Potential changes to future versions of the instrument are discussed.

1. Introduction

Playfulness has been conceptualized in multiple ways. For example, it has been viewed as a component of healthy development (Gordon, 2014); as an attitude that can be modulated (Heimann & Roepstorff, 2018); or as a predisposition (Barnett 2007) or personality trait (Proyer, 2017a) that allows individuals to frame their life experiences in meaningful and constructive ways. Although playfulness is an important construct, it remains understudied, particularly within the adult population (Proyer, 2017b). One explanation for the state of research on playfulness is the definitional tension that exists between play and playfulness. This tension is yet to be resolved, but research in both adult-related play (Davis & Bergen, 2014) and playfulness (Yarnal & Qian, 2011) continue to evolve despite the lack of comprehensive definitions.

Playfulness has been linked to outcomes such as overall health and well-being (Gordon, 2014) but its effects and implications are perhaps most consequential in the realm of education. For children, play and playfulness are the center of their early educational experiences because teachers understand the value of play in the social and emotional development of children (Singer, 2013). There is empirical evidence supporting the academic benefits of play and play-related curricula (Vogt et al., 2018), but most scholars agree that early education should not focus on academic progress at the expense of children's play (Hirsh-Pasek et al., 2009). In terms of adults, playfulness is an important construct because of its contribution to important academic variables like creativity and innovation (Bateson et al., 2013). The importance of playfulness is also evident in Magnuson and Barnett's (2013) cross-sectional study on the relationship between playfulness and stress and coping. They concluded that "playfulness serves a strong adaptive function with university students, providing them with specific cognitive resources from which they can manifest effective coping behaviors in the face of stressful situations" (p. 129).

Research in playfulness has advanced in terms of theoretical foundations and there has also been significant progress in terms of instrumentation. Several instruments have been developed to study playfulness and each focuses on specific conceptions of playfulness. For example, the Playfulness Scale for Young Adults (Barnett, 2007) focused on the general nature of playfulness in young adults while the Adult Playfulness Trait Scale (Shen et al., 2014) considered playfulness as a personality trait. As research in playfulness continues to mature, it is imperative that scholars evaluate the instruments they use because the sensitivity of these instruments will determine the explanatory power of principles and theories that govern playfulness. Thorough evaluations of measurement instruments will ensure that advances in playfulness-related research remain credible and meaningful. The current study addresses this call by using Rasch analysis to evaluate Proyer's (2017a) Other-directed, Lighthearted, Intellectual, and Whimsical Playfulness Scale (OLIW) instrument.

The OLIW instrument measures four traits of adult playfulness: Other-directed (OD), which involves playfulness in social settings; Lighthearted (LH), which reflects spontaneity and improvisation; Intellectual

☆ Conflicts of interest: We have no known conflict of interest to disclose.

* Corresponding author.
E-mail address: davisdr@MiamiOH.edu (D.R. Davis).

https://doi.org/10.1016/j.ijedro.2021.100054
Available online 1 June 2021

(IN), which focuses on playful thoughts or ideas, and Whimsical (WH), which is characterized by a fascination with the odd, strange, or different (Proyer, 2017a). Each construct is measured using seven items and each item is rated using a seven-step Likert scale ranging from "strongly agree" to "strongly disagree" (Appendix A presents the 28-item instrument). The OLIW is a theory-based instrument that was developed and evaluated using a range of statistical techniques including principal component analysis and exploratory and confirmatory factor analyses (Proyer, 2017a; Proyer & Jehle, 2013). It has been used in a number of studies to date, including an exploration of the relationship between playfulness and well-being in adults (Farley et al., 2020) and an examination of adult playfulness and the impostor phenomenon (Brauer & Proyer, 2017). Similarly, the theoretical structure of the OLIW has been used in several studies, including Alatalo's (2018) exploration of the connection between playfulness and well-being in retail jobs and Chick's et. al (2020) work on playfulness and assortative mating. The OLIW has the potential to become an important instrument in research on adult playfulness, and consequently, it must be thoroughly evaluated.

One of the most powerful tools available for evaluating rating scale instruments is Rasch analysis. Rasch analysis provides critical guidance when developing instrumentation, when preparing data for statistical analyses, and when evaluating instruments developed using other methods (Boone et al., 2014). Rasch techniques were developed by Georg Rasch (Rasch, 1960) and expanded by Benjamin Wright (Wright & Stone, 1979). Wright and Masters (1982) provided mathematical details of the model. Although Rasch analysis is mathematically complex, it is useful and accessible to scholars across academic disciplines (Boone & Staver, 2020).

There are many reasons why Rasch methods should be used for analyzing rating scale data. First, Rasch methods provide numerous techniques to evaluate instrument functioning. Second, Rasch provides Person Measures expressed on a linear scale and these measures should be used for parametric statistical tests. Finally, Rasch measurement allows the computation of "item difficulties" that can be used to construct Wright Maps which are subsequently used to evaluate the construct validity of a measurement instrument (Holmefur & Krumlinde-Sundholm, 2016).

The purpose of this study was to evaluate each of the four OLIW constructs using Rasch methods. This appears to be the first time that the OLIW has been evaluated using Rasch methods. This study is a significant addition to the literature because Rasch analysis not only provides insights into the overall functioning of the instrument, but Rasch analysis provides Wright Maps which provide item ordering and spacing that will be invaluable to researchers as they use the OLIW for research on playfulness-related constructs.

2. Method

2.1. Sample and procedure

Data were collected over four semesters from undergraduate students (n = 908) who were enrolled in a human development course at a large mid-western university. The response rate was 75%. In terms of data collection, students provided informed consent and they then completed the OLIW survey using the university's survey system.

2.2. Analyses

We used the Rasch software program Winsteps (Linacre, 2021a) to analyze the data and we used the Rasch Rating Scale Model for our analysis. All text data were re-coded to numbers. For example, "strongly disagree" was re-coded to 1 and "disagree" was re-coded to 2. The Winsteps program reversed the coding of negative items prior to analysis. We investigated the dimensionality of the scales, the overall functioning of the rating scales, varied reliability indices, as well as the Wright Maps generated from the data. The analyses we conducted are commonly used by researchers conducting a Rasch rating scale analysis, but researchers do vary in how they approach a specific analysis goal (e.g. evaluating category probability curves). We used past studies to guide our analyses and we believe that these analyses provide an informative and expanded evaluation of the OLIW instrument. Our analyses provide information that can be used to inform research on adult playfulness.

2.2.1. Evaluating unidimensionality

Unidimensionality is a requirement in Rasch analysis (Wright & Stone, 1979). Evidence of unidimensionality supports using a set of items to determine where a respondent falls along a construct. We evaluated unidimensionality using Fit statistics, Principal Component Analysis of Residuals (PCAR), and Point-Measure Correlations.

When the data are determined to "fit" the Rasch model, it means that there is evidence of unidimensionality (Brown et al., 2016; Li et al., 2016). To explore fit, we used the Rasch fit statistics Infit Mean-Square (MNSQ) and Outfit MNSQ. The Infit statistic is a data inlier-sensitive statistic, while the outfit statistic is a data outlier-sensitive statistic. Following O'Connor et al. (2016), we used an MNSQ range of 0.5 to 1.5 to indicate acceptable fit. The Winsteps manual (Linacre, 2021b) states that MNSQ values between 0.5 and 1.5 are "Productive of measurement" (p. 378). It is important to note that Rasch researchers have used several different MNSQ ranges, for example, Veas et al. (2016) used the 0.8 to 1.2 range while Moeini et al. (2016) considered 0.6 to 1.4 to be the optimal range for MNSQs. If we used some of the other fit ranges, some of our dataset items would be flagged for misfit (those items outside the acceptable range) or some of our dataset items would be classified as borderline.

PCAR is another technique used to investigate dimensionality (de Haan et al., 2011; Dougherty et al., 2011). This technique investigates whether the level of noise seen in a dataset is above what would be expected by random noise. When the level of noise is above the level of predicted random noise, it indicates the possible presence of confounding variables. Bradley et al. (2016) suggested that when using PCAR, there is evidence of unidimensionality when the value of the unexplained variance in the first contrast is below 2.0 eigenvalue units.

Finally, dimensionality can be evaluated using point-measure correlations. Li et al. (2016) recommended a benchmark of 0.3 to establish unidimensionality and viewed items with point-measure correlations above 0.3 as acceptable.

2.2.2. Evaluating rating scale functioning

It is important to evaluate the manner in which the rating scale of an instrument functions to determine if the rating scale abides by the requirements of measurement. For example, it is important to know if respondents are using all the rating scale categories. Also, it is important to evaluate if there is evidence that rating scale categories each contribute to providing measurement information regarding respondents. We used the Linacre (1999, 2004) guidelines to investigate the rating scale functioning of the four OLIW constructs. These guidelines have been used by researchers using Rasch techniques (examples include Marakis et al., 2016; O'Connor et al., 2016; and Finger et al., 2012). The guidelines are as follows:

1 There must be at least 10 observations of each category.
2 Observations should be uniformly distributed. According to O'Connor et al. (2016), observations should be uniform across the peaks of frequently used categories and there should be a gradual flattening for less frequently used categories.
3 Average measures should advance monotonically with each category.
4 The outfit MNSQ should be less than 2.0. Linacre (2002) noted that levels higher than 2.0 indicate unexplained randomness which may distort the measurement system. Ideally, values would be between 0.5 and 1.5.

5 Step calibrations should advance monotonically.
6 There should be a reciprocal relationship between the meanings of ratings and measure.
7 Gaps between categories should be larger than 1.4 logits.
8 Gaps between categories should be less than 5.0 logits.

2.2.3. Evaluating person and item reliability and person and item separation

Rasch analysis provides the tools to assess person and item reliability, as well as person and item separation for survey scales (Eggert et al., 2017). These statistics help provide an indication of how well the items are able to distribute the respondents along the construct and how well the respondents are able to separate the items used to define the construct. Person and item reliability are analogous to Cronbach alpha, where a higher value indicates greater reliability. Malec et al. (2007) suggested critical values of .80 for person reliability, .90 for item reliability, 2.0 for person separation, and 4.0 for item separation.

2.2.4. Evaluating targeting, floor, and ceiling effects

Targeting, floor effects, and ceiling effects are important measures to consider in Rasch analysis (Khadka et al., 2016; Wong et al., 2018). Targeting can be evaluated by looking at the difference between the average person measure and the average item measure of the dataset. Finger et al. (2012) suggested that a difference below 1.00 logits represents good targeting of items and persons. Floor and ceiling effects concern the presence of respondents who are at the highest and lowest ends of the scale. Fisher (2007) categorized an instrument as "poor" when floor and ceiling effects were greater than 5% and "excellent" when the effects were less than 0.5%.

2.2.5. Evaluating instrument functioning and construct validity

A Wright Map is a powerful Rasch tool that is used to evaluate the functioning of an instrument. It allows researchers to review the ordering and spacing of items in an effort to find patterns that reveal the structure of the instrument. If item ordering and spacing match what is predicted from theory, that match provides evidence of instrument construct validity (Boone & Staver, 2020). A Wright Map typically presents survey items on the right side of a vertical line and person measures on the left side. In our Wright Maps, the items are organized from "easier to agree with or endorse" (at the bottom) to "more difficult to agree with or endorse" (at the top). When reviewing a Wright Map, it is important to note the location of items. Overlapping items (items at the same location on the Wright Map) indicate redundancy for those items in terms of marking the trait and gaps between items indicate potential weakness in the instrument (Chen et al., 2017). In the interest of clarity, only the survey items are included in our Wright Maps.

3. Results

This section reports the results of the Rasch analysis of the Other-directed (OD), Lighthearted (LH), Intelligence (IN), and Whimsical (WH) constructs in the OLIW instrument.

3.1. Unidimensionality

Our overall evaluation of the OLIW instrument supports the assertion that each of the four item-sets measures a single construct or trait. Table 1 presents a summary of fit analysis of the OLIW instrument. Of importance is that all items in the OD, LH, IN, and WH constructs fall within the 0.5 to 1.5 range of MNSQ for both infit and outfit. The result of the fit analysis suggests that each of the four item-sets measures a single construct. However, there are a small number of items that are close to misfitting. Similarly, the PCAR results were also positive in that the scales exhibited an unexplained variance in the first contrast below 2.0 eigenvalue units. The PCAR result provided additional evidence that the scales are unidimensional. Table 1 shows that all items in each construct had point-measure correlations values greater than 0.3. This provides further evidence of the unidimensionality of the four OLIW constructs.

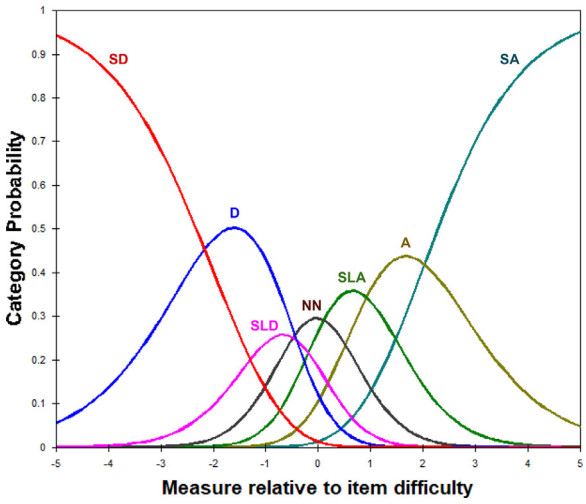

Fig. 1. Category Probability Curve for the Whimsical Construct (WH). Note. SD = strongly disagree, D = disagree, SLD = slightly disagree, NN = neither agree nor disagree, SLA = slightly agree, A = agree, and SA = strongly agree.

3.2. Rating scale functioning

Our evaluation found that most of the rating scales used in the OLIW abided by the suggested guidelines for a well-functioning rating scale. This means that in general, the rating scales were performing as expected.

Table 2 provides the results from our evaluation of the functioning of the rating scale used in the OLIW. Focusing on the Linacre (1999, 2004) guidelines, we see that the first guideline is satisfied for all rating scales, as evident in the "Observed Counts" column in Table 2 where there are at least 10 observations of each category. The second guideline requires that observations be uniformly distributed and it is evaluated by reviewing the distribution of responses in the "Observed Count" column in Table 2. This guideline is generally satisfied for all rating scales. For the third guideline, average measures should advance monotonically with the categories. This guideline is satisfied for all rating scales as seen in the "Observed Average" column in Table 2 where, for example, the WH measure increases monotonically from a low of -0.82 to a high of 1.00 logits. Next, guideline four requires that outfit MNSQ values are less than 2.0 and this requirement is satisfied for all scales. The "Outfit MNSQ" column in Table 2 shows that all average MNSQ values are below 2.0.

For the fifth guideline, we should see monotonically advancing step calibrations. The "Andrich Thresholds" column in Table 2 shows that this criterion was satisfied except for the small deviations in the LH, IN, and WH rating scales. We also examined the probability curve for each rating scale. Ideally, categories should have individual peaks cutting across each other at the step calibration or threshold measure. Fig. 1 shows the probability curve for the rating scale used by the WH construct and Appendix C provides the probability curves for the rating scales used in the OD, LH, and IN constructs. Looking at Fig. 1, we see that each curve, except "slightly disagree," has an individual peak which is most probable for some portion of the horizontal axis. In essence, the "slightly disagree" curve for the WH construct is completely covered by other curves. We see a similar result for the probability curves for the OD construct. The probability curves for the LH and IN constructs were similar in that the "neither agree nor disagree" curves for both the LH and IN constructs were completely covered by other curves. Overall,

Table 1
Summary of the Rasch item statistics for each construct of the OLIW instrument.

Construct and Item Number	Total Score	Total Count	Measure (Logits)	Model S.E. (Logits)	Infit MNSQ	Outfit MNSQ	P. M. Corr.
Other-directed Construct (OD)							
OD 03	5362	898	-0.49	0.04	0.72	0.71	0.54
OD 07	4279	898	0.63	0.03	1.12	1.09	0.57
OD 11	5548	897	-0.80	0.04	1.37	1.14	0.51
OD 15*	4072	897	0.79	0.03	0.96	0.98	0.59
OD 19	5017	896	-0.06	0.03	0.96	0.99	0.57
OD 23	5009	895	-0.06	0.03	0.86	0.96	0.54
OD 27*	4969	895	-0.02	0.03	1.48	1.37	0.54
Lighthearted Construct (LH)							
LH 02*	3247	908	0.43	0.03	1.11	1.15	0.50
LH 06	3421	908	0.32	0.03	1.15	1.18	0.57
LH 10	4734	908	-0.62	0.03	1.13	1.16	0.46
LH 14	4556	907	-0.48	0.03	0.74	0.77	0.63
LH 18	3802	907	0.06	0.03	0.83	0.84	0.61
LH 22	4096	905	-0.14	0.03	1.06	1.09	0.46
LH 26	3237	905	0.43	0.03	1.00	1.02	0.65
Intellectual Construct (IN)							
IN 01	4393	908	-0.41	0.03	1.13	1.17	0.32
IN 05*	3936	908	-0.06	0.03	1.09	1.09	0.45
IN 09	4066	908	-0.15	0.03	0.72	0.73	0.56
IN 13*	3375	907	0.35	0.03	0.82	0.83	0.52
IN 17	3830	907	0.02	0.03	0.93	0.93	0.58
IN 21*	3771	905	0.05	0.03	0.98	1.00	0.50
IN 25	3562	905	0.21	0.03	1.32	1.35	0.46
Whimsical Construct (WH)							
WH 04	3909	908	-0.10	0.03	0.82	0.86	0.56
WH 08	3207	908	0.44	0.03	1.05	1.04	0.65
WH 12	4226	907	-0.36	0.03	1.08	1.10	0.53
WH 16	3981	907	-0.16	0.03	1.03	1.04	0.53
WH 20	3791	906	-0.02	0.03	0.77	0.79	0.66
WH 24	3876	905	-0.09	0.03	1.15	1.15	0.58
WH 28	3381	905	0.30	0.03	1.12	1.14	0.47

Note. * indicates reversed items. Total score is the total raw score of all respondents who answered the item. Total count is the total number of respondents who answered the item. Measure is the Rasch item measure in logit units. Model S.E. is the standard error of the item measure in logit units. Infit MNSQ is an inlier-sensitive fit statistic. Outfit MNSQ is an outlier-sensitive fit statistic. P. M. Corr is the Pearson product-moment correlation coefficient. Please note that the meaning of a logit is unique to each construct because each construct represents a different variable.

these results indicate that some rating scale categories might need to be combined.

Guideline six evaluates "Coherence" which refers to the relationship between measures and categories, that is, do measures imply categories (M->C) and do categories imply measures (C->M). Typically, coherence results greater than 40% are acceptable (Linacre, 1999), but low values do not necessarily indicate problematic categories. For the OLIW, (M->C) was not successful given that only the WH construct had a majority of the rating scale categories (four of seven) reach the 40% threshold. Similarly, (C->M) was unsuccessful in that the best functioning constructs, OD and IN, only had two (of seven) rating scale categories meet the guideline.

Finally, to examine guidelines seven and eight we use the values in the "Andrich Thresholds" column in Table 2 to compute the distance between adjacent Andrich Thresholds (these are the gaps). The results are sub-optimal for guideline seven where no rating scale had more than two adjacent categories with gaps greater than 1.4 logits. However, guideline eight was fully satisfied as all adjacent categories in all rating scales had gaps less than 5.0 logits.

3.3. Person and item reliability and person and item separation

Table 3 summarizes the person reliability, person separation, item reliability, and item separation of the OLIW constructs. Results were very positive for item reliability and item separation, where values were above 0.90 and 4.0 respectively. However, the constructs performed sub-optimally with respect to person reliability and person separation, where values did not meet the critical values of 0.80 for person reliability and 2.0 for person separation.

3.4. Targeting, floor, and ceiling effects

Three of the four OLIW constructs demonstrated excellent targeting, no floor effect, and no ceiling effect. Looking at Table 4, we see that with the exception of OD, all the values in the "Difference" column were less than 1.00 logits, indicating good targeting of items and persons. Similarly, except for OD, the values in the "% Floor" and "% Ceiling" columns in Table 4 were all less than 0.5% and therefore we can consider them excellent.

3.5. Instrument functioning and construct validity

We also used Wright Maps to evaluate the functioning of the OLIW constructs. Fig. 2 presents the Wright Map for the Whimsical (WH) construct and Appendix B presents the remaining constructs. What we notice in Fig. 2 is that there is an item gap (between WH 28 and WH 20) and a clear grouping of items (WH 20, WH 24, WH 4, and WH 16). We see a similar pattern with the OD construct. In some instances, such as with items OD 19 and OD 23, the grouping includes items that are located in or near the same position. The Wright Maps for the LH, and IN constructs all show at least one gap between items, but they do not have pronounced groupings (overlaps) of items as seen in the Wright Maps for the OD and WH constructs. The Wright Maps show a good marking of each construct by the items despite the presence of some gaps and some grouping of items. However, through the addition of items to the survey there could be an optimization of item distribution, this is especially the case for the OD and WH constructs. We consider this in the discussion section of this paper.

Table 2
Summary of the rating scale performance the OLIW Instrument.

Category Label	Observed Count	Observed %	Observed Average (Logits)	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Gap (Logits)
Other-directed Construct (OD)							
Strongly Disagree	56	1	0.05	1.68	1.91	None	
Disagree	246	4	0.07	1.40	1.56	-1.76	1.33
Slightly Disagree	370	6	0.18	1.11	1.12	-0.43	0.08
Neither Agree or Disagree	701	11	0.33	0.90	0.90	-0.35	0.45
Slightly Agree	1205	19	0.74	0.96	0.84	0.10	0.40
Agree	2117	34	1.32	0.94	0.91	0.50	1.44
Strongly Agree	1581	25	2.05	0.89	0.94	1.94	
Lighthearted Construct (LH)							
Strongly Disagree	306	5	-0.78	0.94	1.02	None	
Disagree	828	13	-0.43	1.07	1.12	-1.58	0.98
Slightly Disagree	1070	17	-0.25	0.90	0.88	-0.60	0.66
Neither Agree or Disagree	904	14	0.06	0.99	1.04	0.06	0.45
Slightly Agree	1552	24	0.28	0.99	0.99	-0.39	1.03
Agree	1271	20	0.57	1.04	1.05	0.64	1.24
Strongly Agree	417	7	0.96	1.05	1.08	1.88	
Intelligence Construct (IN)							
Strongly Disagree	124	2	-0.47	1.03	1.06	None	
Disagree	732	12	-0.29	1.01	1.04	-2.15	1.43
Slightly Disagree	1232	19	-0.12	0.97	1.00	-0.72	0.70
Neither Agree or Disagree	1221	19	0.07	0.90	0.90	-0.02	0.15
Slightly Agree	1690	27	0.25	0.97	0.98	-0.17	0.94
Agree	1128	18	0.51	0.98	1.00	0.77	1.53
Strongly Agree	221	3	0.81	1.15	1.11	2.30	
Whimsical Construct (WH)							
Strongly Disagree	193	3	-0.82	1.09	1.14	None	
Disagree	904	14	-0.44	1.03	1.06	-2.17	1.76
Slightly Disagree	984	16	-0.19	0.99	0.99	-0.41	0.07
Neither Agree or Disagree	1461	23	0.01	0.92	0.94	-0.48	0.59
Slightly Agree	1546	24	0.28	1.01	1.01	0.11	0.81
Agree	962	15	0.68	0.91	0.93	0.92	1.11
Strongly Agree	296	5	1.00	1.09	1.07	2.03	

Table 3
Summary of the person reliability, person separation, item reliability, and item separation of the OLIW constructs.

OLIW Construct	Person Reliability	Person Separation	Item Reliability	Item Separation
Other-directed (OD)	0.73	1.66	1.0	15.22
Lighthearted (LH)	0.70	1.52	1.0	14.84
Intelligence (IN)	0.60	1.22	0.99	8.27
Whimsical (WH)	0.74	1.67	0.99	9.05

Table 4
A summary of the targeting, floor effect, and ceiling effect observed for each OLIW construct.

OLIW Construct	Average Person Measure (Logits)	Average Item Measure (Logits)	Difference (Logits)	% Floor	% Ceiling
Other-directed (OD)	1.15	0.00	1.15	0.00	0.66
Lighthearted (LH)	0.12	0.00	0.12	0.00	0.00
Intelligence (IN)	0.13	0.00	0.13	0.00	0.00
Whimsical (WH)	0.11	0.00	0.11	0.00	0.00

4. Discussion

We conducted numerous Rasch analyses to investigate the psychometric properties of the four OLIW constructs. The evaluation of unidimensionality using analysis of fit, PCAR, and point-measure correlation suggested that each of the four OLIW constructs only measured one trait. The unidimensionality finding is consistent with Proyer's (2017a) assessment of the instrument's factorial validity, but this finding alone does not provide evidence that the constructs are theoretically sound. The finding of unidimensionality in this study simply confirms that we have found that the constructs are sufficiently unidimensional for meaningful measurement. Our evaluation of targeting and floor and ceiling effects were also positive. The distribution of items along the four constructs was such that few or no respondents had a maximum or minimum measure. This result provides evidence of the instrument's ability to accommodate a wide range of respondent views. Furthermore, the average person measure for each construct was close to the average item measure indicating acceptable targeting.

The reliability and separation indices were reviewed for each construct with respect to persons and items. The item reliability and item separation indices fell within suggested bounds. However, the person reliability and person separation indices were below values suggested by authors such as Malec et al. (2007). There can be many reasons for low person reliability and low person separation, but a common reason is insufficient items for part of a particular construct. One solution that might increase the person reliability and person separation indices is to add items to the constructs. In particular, to fill the gaps observed in the Wright Maps. Also, surveying a student sample with a wider range of attitudes may result in a higher person reliability and higher person separation. We suggest that researchers who use OLIW should monitor these two indices when additional samples of respondents are surveyed

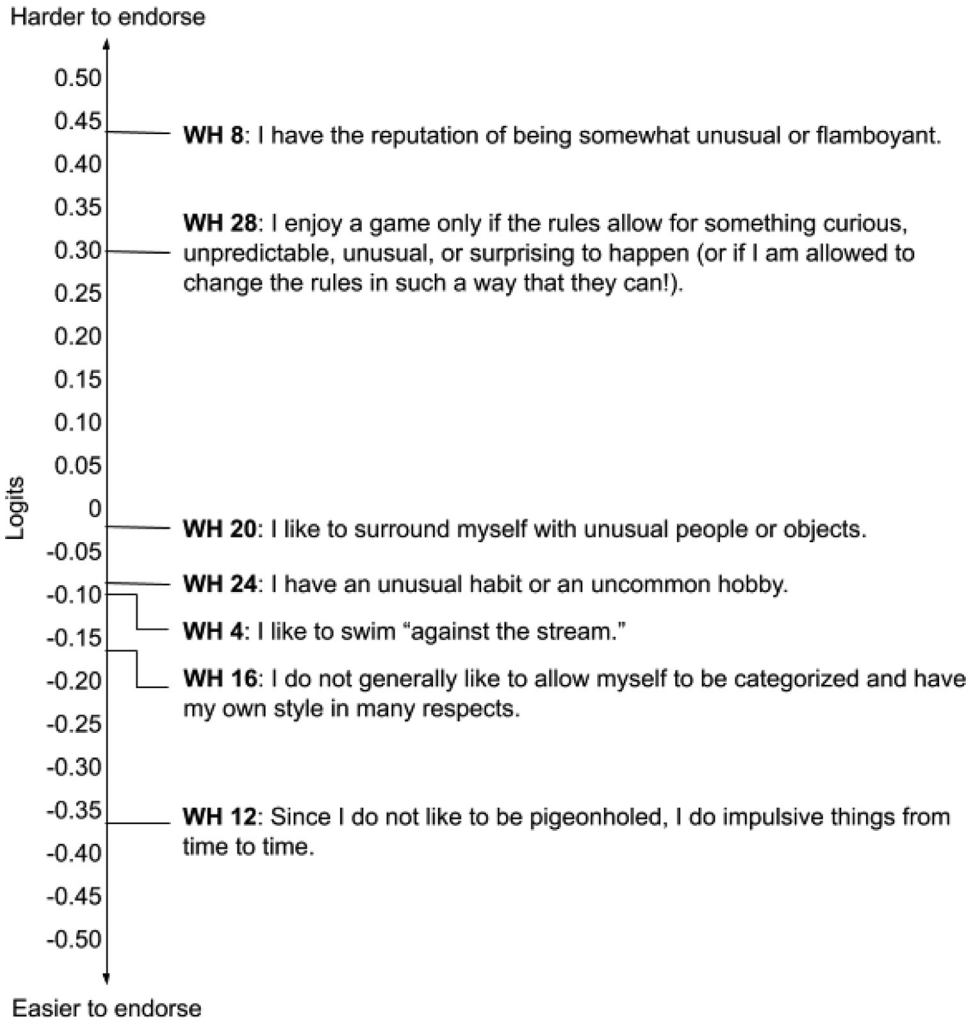

Fig. 2. Wright map for the Whimsical (WH) construct.

with the current OLIW and/or when new items are authored for the instrument.

We used the Wright Map of each construct to examine the spacing and ordering of items. Although the items were generally well-spaced across all the constructs, there were instances where items were close together. For example, the proximity of items LH 2 and LH 26 in Appendix B indicates that those items may be redundant from a measurement perspective (these items mark similar parts of the constructs). In this case, item LH 2 or LH 26 could be deleted to make the instrument more efficiently measure. Similarly, there are instances of gaps between items, for example, the gap between items WH 20 and WH 28 in Fig. 2 could be filled with a new item. This procedure of removing redundant items and filling gaps with items would improve the measurement function of the instrument without increasing the number of items on the instrument. It will be important for future versions of the OLIW to add new items to fill the gaps between separated items and thereby increase the measurement precision of the overall instrument. Our view is that the authoring of new items to fill the item gaps on the Wright Maps is the highest priority for a revised OLIW.

One potential area of concern with the OLIW involves the rating scale categories. Although the rating scales generally functioned in a desirable manner, there is evidence to suggest that combining some adjacent scale categories may be warranted (see Finger et al., 2012 for more on this technique). For example, the probability curve for the WH construct in Fig. 1 shows that it might be useful to combine the two adjacent "agree" categories and also combine the two adjacent "disagree" categories. Given some differences in the performance of the rating scale as a function of construct, it is worth considering that each construct in the OLIW instrument could be rated using its own rating scale. However, one disadvantage of using different rating scales within the same instrument is that the shift in rating scales might confuse some respondents.

As part of our analysis, we experimented with the combining of selected adjacent categories. One experiment involved combining the "disagree" and "slightly disagree" categories as well as combining the "agree" and "slightly agree" categories. Another experiment involved combining the "slightly agree," "neither agree nor disagree," and "slightly disagree" categories into a single central category. In both of those experiments and for all four OLIW constructs, there was an improvement in the location of the category curves, but there was also degradation in some of the other indices that are considered when evaluating a measurement device. For example, in all our experiments with combining scales, there was a lessening in the person reliability values for the constructs. As the functioning of an instrument is always a balancing act with many factors to consider, we felt that in the end, it was best to retain the original rating scale used for each of the OLIW constructs. We feel that it is prudent to use one single rating scale if data for all four constructs are collected at one time from respondents. However, it might be useful to consider some slight alterations in the rating scale if researchers would collect data for just one construct of the OLIW instrument.

Patterns of item-ordering observed on each Rasch Wright Map provides new details regarding each construct. More specifically, the item hierarchy helps define different parts of each construct. For example,

the Wright Map for LH shows examples of Lightheartedness that are exhibited by many people (bottom of the map) and examples of Lightheartedness that are only exhibited by the most lighthearted respondents (top of the map). Proyer (2017a) provided a rationale as to why items were placed into each construct, however, our Wright Maps identify the hierarchy of items within each of the construct. We believe that this result contributes to a deeper understanding of each construct and the overall instrument. It is important to have evidence that a survey item helps to measure a specific construct, but it is equally important to know what part of a construct is being measured by the survey item. For example, the Wright Map for the WH construct (Fig. 2) shows that item WH 12 "Since I do not like being pigeonholed, I do impulsive things from time to time" was located in a much different part of the construct in comparison to item WH 8 "I have the reputation of being somewhat unusual or flamboyant." Both items define the WH construct, but they define different parts of the construct. We suggest that by our Wright Maps providing the important information of the location of each item along each construct, theories related to adult playfulness can be further detailed.

In our study to inform research on adult playfulness and our teaching of the topic, we were interested in the overall Wright Maps for the four OLIW constructs. For those researchers who wish to make comparisons between subgroups of participants (e.g. male respondents to female respondents), it will be important to consider conducting a Differential Item Functioning (DIF) analysis (Costa, et al., 2016). A DIF analysis helps one to evaluate the extent to which the instrument functions for comparison groups. If a DIF analysis reveals that there are items which define a trait in a different manner for different comparison groups, then there are a wide range of steps which can then be taken to address the presence of DIF. In our study, we evaluated DIF with respect to gender and we did not find evidence of DIF as a function of gender.

5. Conclusion

Overall, we feel that the four constructs of the OLIW exhibit acceptable psychometric properties as viewed from a Rasch measurement perspective. Each of the OLIW's four constructs appeared to define a single trait and the measurement scales generally functioned as expected. Item reliability and item separation were acceptable for each construct, but person reliability and person separation did not reach targeted values. It is important to collect additional data from varied populations to further investigate the factors which impacted person reliability and person separation. Wright Maps indicated that items were generally well-spaced, although there were instances of clustering of items and also gaps between items. Proyer (2017a) provided a rationale for the assignment of each item to its respective construct, but the author did not define the location of each item along its construct. Our Wright Maps provide this additional important information.

For future versions of the instrument, we suggest adding items to each construct in an effort to fill the few gaps present between items on the Wright Maps. Similarly, the overall efficiency of the instrument could be improved by removing items that mark the same or similar portion of the construct. Also, future researchers might consider changing the rating scale categories for the OLIW. However, having multiple rating scales on the same instrument could be confusing for respondents. We also suggest that future researchers using the OLIW instrument compute Rasch person measures and use these measures for parametric statistical tests instead of the typical mean values of the coded responses. This is because raw score totals are non-linear. Lastly, the issue of item-ordering is important because the ordering should reflect theories of adult playfulness. We suggest that as theories of adult playfulness are further refined, it is important to compare those theories to what is revealed in our Wright Maps. This would allow theories to be confirmed or refined as needed. It is also important to collect data from different populations to see if the patterns of the Wright Maps are stable across factors such as age or cultural identity.

Declaration of Competing Interest

None.

Appendix A

Table A1

Table A1
The 28 items in the OLIW instrument.

Construct	Number	Item
IN	1	In the final account, a discussion is nothing other than playing with -and an exchange of ideas.
LH	2*	I do not live from day to day at all; I rather plan ahead long in advance.
OD	3	I can use my playfulness to do something nice for other people, or to cheer them up.
WH	4	I like to swim "against the stream."
IN	5*	I do not like tasks where you have to try a few things out and have to puzzle something out, before arriving at a good solution
LH	6	I don't worry about most of the things that I have to do, because there will always be some kind of a solution.
OD	7	Also as an adult I still like to play good natured, funny tricks on others; to play small good-natured pranks on others.
WH	8	I have the reputation of being somewhat unusual or flamboyant.
IN	9	If I want to develop a new idea further and think about it, I like to do this a playful manner.
LH	10	I am a lighthearted person.
OD	11	I have close friends with whom I can just fool around and be silly.
WH	12	Since I do not like being pigeonholed, I do impulsive things from time to time.
IN	13*	When thinking about a problem, I look for a fixed scheme for the solution and only rarely rely on a playful approach to solve the problem.
LH	14	Many people take their lives too seriously; when things don't work you just have to improvise.
OD	15*	It only distracts me from work if my colleagues at work are fooling around and want to involve me. There is no use in this for me and I do not even participate in such things.
WH	16	I do not generally like to allow myself to be categorized and have my own style in many respects.
IN	17	If I have to learn something new under time pressure, I try to find a playful approach to the topics—this helps me learn.
LH	18	"Wait and see" is often a better approach than spending much time pondering.
OD	19	I enjoy re-enacting things with close friends that we have experienced together (e.g., a funny incident that we like to remember).
WH	20	I like to surround myself with unusual people or objects.
IN	21*	If one has a concrete task to perform, there is no room for playfulness. This only detracts from the work.
LH	22	It happens sometimes (at work or in leisure time) that I do something and do not really think about the possible consequences and all the things that could be happening.
OD	23	I can express my feelings towards my romantic partner in a playful way.
WH	24	I have an unusual habit or an uncommon hobby.
IN	25	I can always think of something to do and I am never bored.
LH	26	If I am free to choose, I prefer to work somewhat chaotic and unplanned than planning everything up to the smallest detail.
OD	27*	I do not at all enjoy parlor games (e.g., card games, or a games of dice) and engaging in a playful way with other people.
WH	28	I enjoy a game only if the rules allow for something curious, unpredictable, unusual, or surprising to happen (or if I am allowed to change the rules in such a way that they can!).

Note. * indicates reversed items.

Appendix B

Figures B1, B2, B3

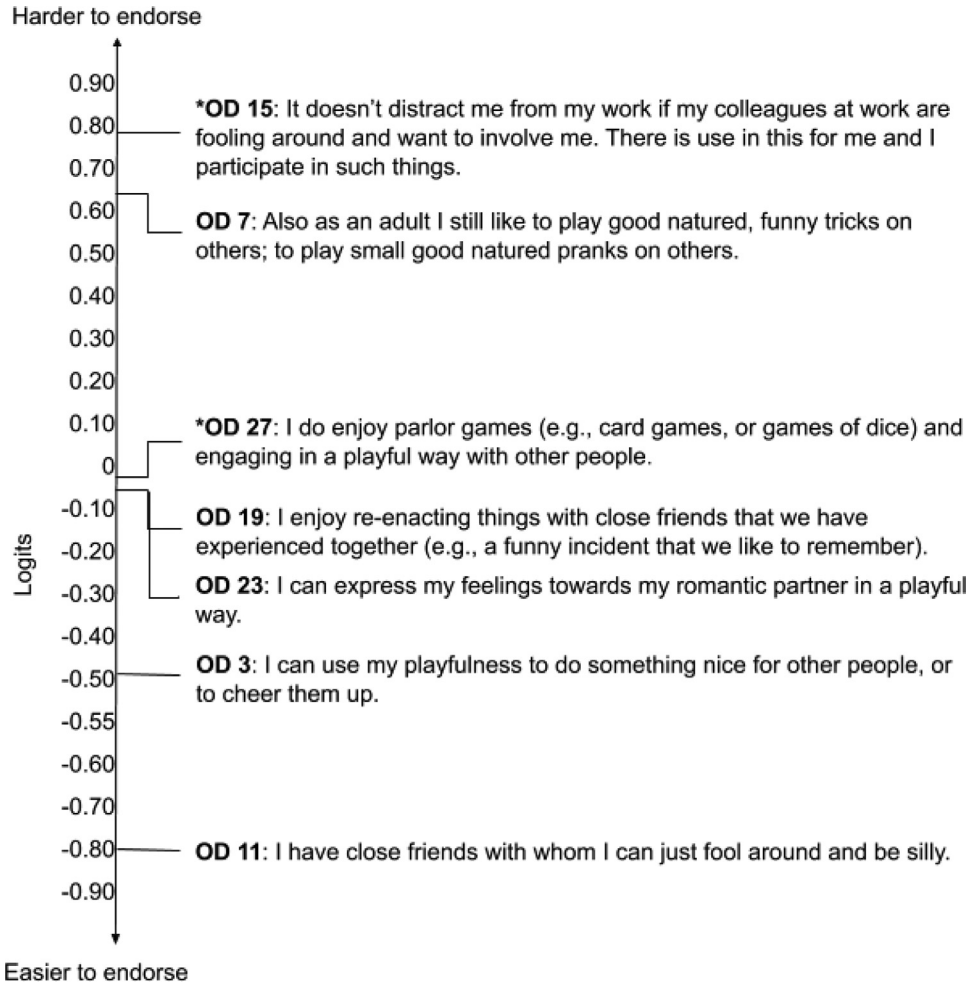

Fig. B1. Wright Map for the Other-directed Construct (OD). Note. * indicates that the item was negatively worded in the OLIW, but presented with non-negative wording in the Wright Map.

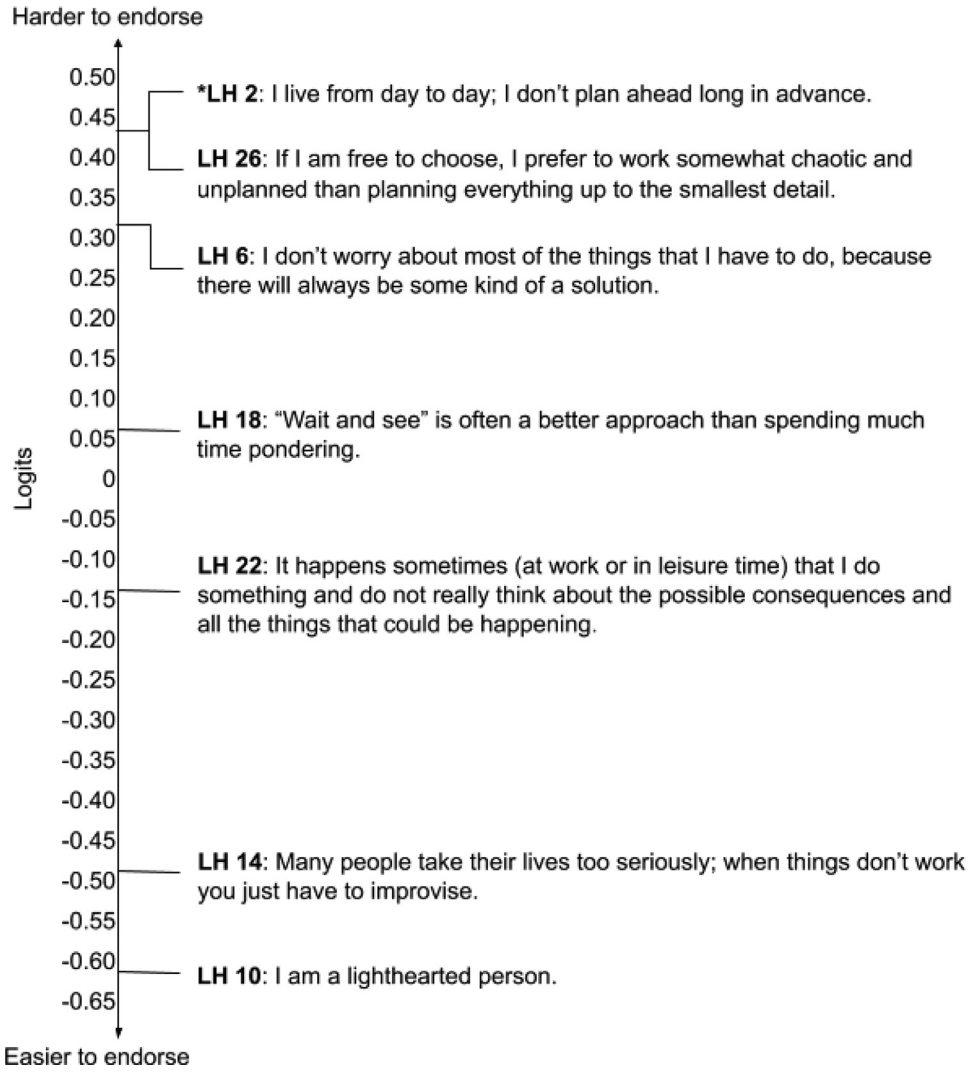

Fig. B2. Wright Map for the Lighthearted Construct (LH). Note. * indicates that the item was negatively worded in the OLIW, but presented with non-negative wording in the Wright Map.

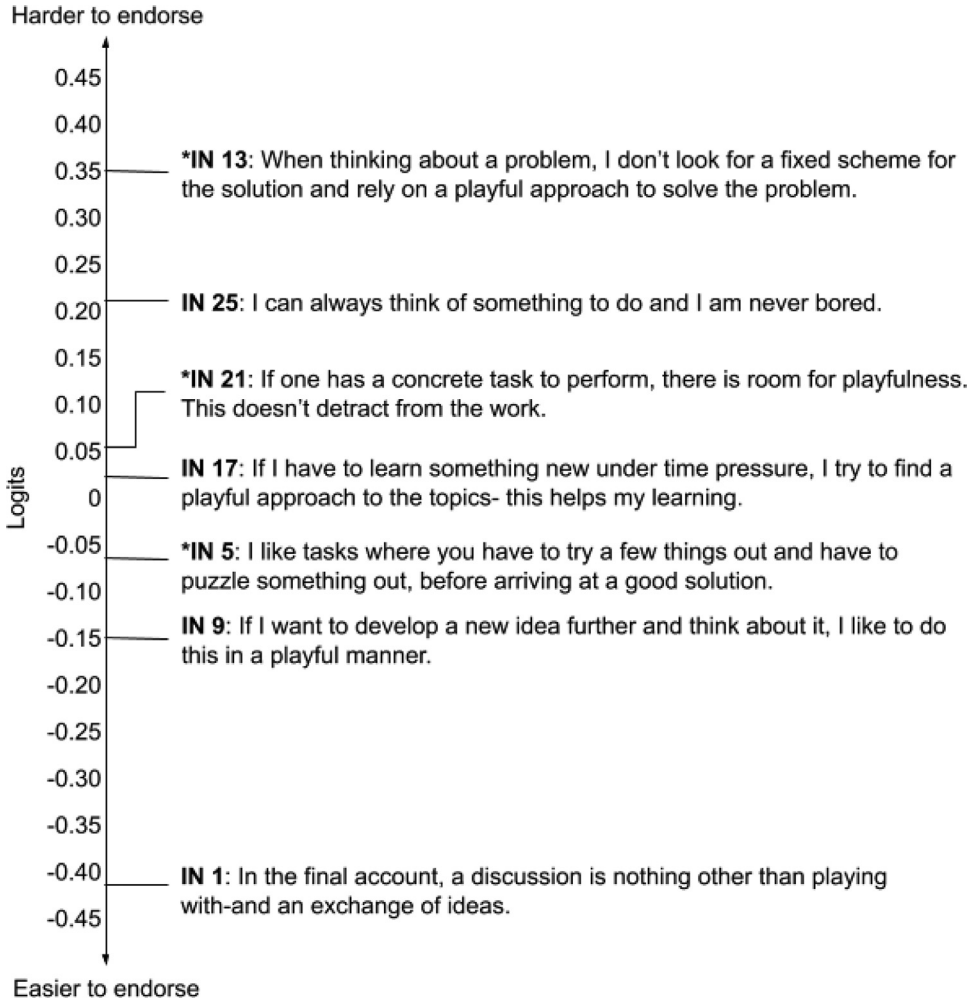

Fig. B3. Wright Map for the Intellectual Construct (IN). Note. * indicates that the item was negatively worded in the OLIW, but presented with non-negative wording in the Wright Map.

Appendix C

Figs. C1, C2, C3

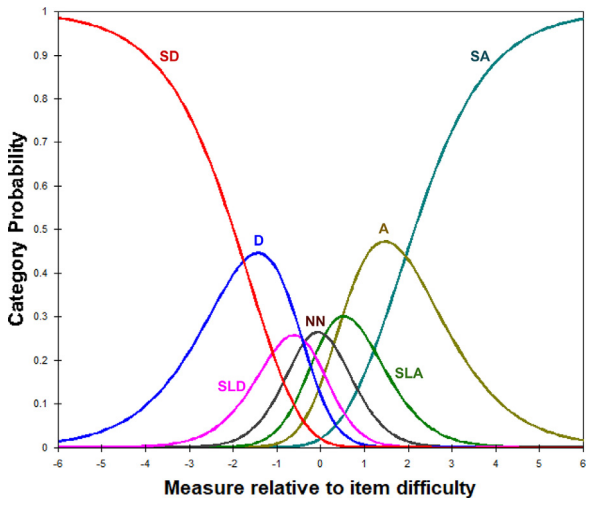

Fig. C1. Category Probability Curve for the Other-directed Construct (OD). Note. SD = strongly disagree, D = disagree, SLD = slightly disagree, NN = neither agree nor disagree, SLA = slightly agree, A = agree, and SA = strongly agree.

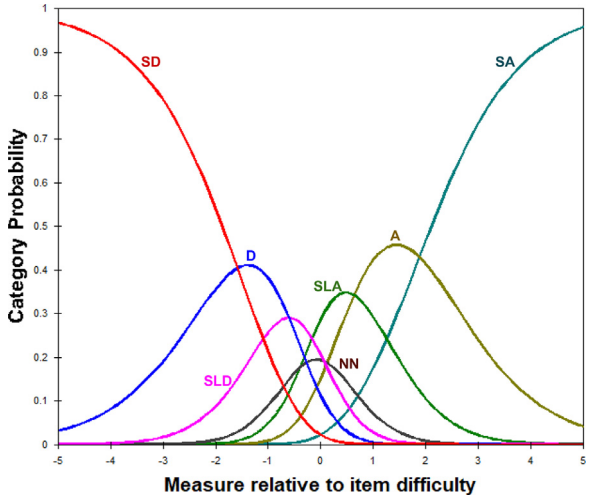

Fig. C2. Category Probability Curve for the Lighthearted Construct (LH). Note. SD = strongly disagree, D = disagree, SLD = slightly disagree, NN = neither agree nor disagree, SLA = slightly agree, A = agree, and SA = strongly agree.

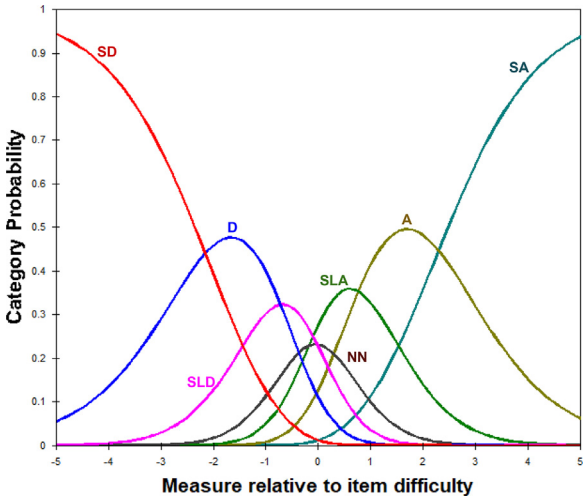

Fig. C3. Category Probability Curve for the Intellectual Construct (IN). Note. SD = strongly disagree, D = disagree, SLD = slightly disagree, NN = neither agree nor disagree, SLA = slightly agree, A = agree, and SA = strongly agree.

References

Alatalo, S., Oikarinen, E. L., Reiman, A., Tan, T. M., Heikka, E. L., Hurmelinna-Laukkanen, P., Muhos, M., & Vuorela, T. (2018). Linking concepts of playfulness and well-being at work in retail sector. *Journal of Retailing and Consumer Services, 43*, 226–233.

Barnett, L. A. (2007). The nature of playfulness in young adults. *Personality and Individual Differences, 43*(4), 949–958. 10.1016/j.paid.2007.02.018.

Bateson, P., Bateson, P. P. G., & Martin, P (2013). *Play, playfulness, creativity and innovation*. Cambridge University Press.

Boone, W., Staver, J., & Yale, M. (2014). *Rasch analysis in the human sciences*. Springer Publishers.

Boone, W., & Staver, J. (2020). *Advances in Rasch analysis in the human sciences*. Springer Publishers.

Brauer, K., & Proyer, R. T. (2017). Are Impostors playful? *Testing the association of adult playfulness with the impostor phenomenon.* Personality and Individual Differences, *116*, 57–62.

Brown, R. L., Obasi, C. N., & Barrett, B. (2016). Rasch analysis of the WURSS-22 dimensional validation and assessment of invariance. *Journal of Lung, Pulmonary, and Respiratory Research, 3*(2), 76. 10.15406/jlprr.2015.03.00076.

Chen, S., Zhu, X., & Kang, M. (2017). Development and validation of an energy-balance knowledge test for fourth-and fifth-grade students. *Journal of Sports Sciences, 35*(10), 1004–1011.

Chick, G., Proyer, R., Purrington, A., & Yarnal, C. (2020). Do birds of a playful feather flock together? Playfulness and assortative mating. *American Journal of Play, 12*(2), 178–215.

Costa, A. B., de Lara Machado, W., Ruschel Bandeira, D., & Nardi, H. C (2016). Validation study of the revised version of the Scale of Prejudice Against Sexual and Gender Diversity in Brazil. *Journal of Homosexuality, 63*(11), 1446–1463. 10.1080/00918369.2016.1222829.

de Haan, J., Schep, N., Tuinebreijer, W., Patka, P., & den Hartog, D. (2011). Rasch analysis of the Dutch version of the Oxford elbow score. *Patient Related Outcome Measures, 2*, 145–149.

Davis, D. R., & Bergen, D. (2014). Relationships among play behaviors reported by college students and their responses to moral issues: A pilot study. *Journal of Research in Childhood Education, 28*(4), 484–498.

Dougherty, B. E., Nichols, J. J., & Nichols, K. K. (2011). Rasch analysis of the Ocular Surface Disease Index (OSDI) (2011). *Investigative Ophthalmology & Visual Science, 52*(12), 8630–8635. 10.1167/iovs.11-8027.

Eggert, S., Nitsch, A., Boone, W. J., Nückles, M., & Bögeholz, S. (2017). Supporting students' learning and socioscientific reasoning about climate change—the effect of computer-based concept mapping scaffolds. *Research in Science Education, 47*(1), 137–159. 10.1007/s11165-015-9493-7.

Farley, A., Kennedy-Behr, A., & Brown, T. (2020). An investigation into the relationship between playfulness and well-being in Australian adults: An exploratory study. *OTJR: Occupation, Participation and Health, 41*(1), 56–64. 10.1177/1539449220945311.

Finger, R. P., Fenwick, E., Pesudovs, K., Marella, M., Lamoureux, E. L., & Holz, F. (2012). Rasch analysis reveals problems with multiplicative scoring in the Macular Disease Quality of Life questionnaire. *Ophthalmology, 119*(11), 2351–2357. 10.1016/j.ophtha.2012.05.031.

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Gordon, G. (2014). Well played: The origins and future of playfulness. *American Journal of Play, 6*(2), 234–266.

Heimann, K. S., & Roepstorff, A. (2018). How playfulness motivates–Putative looping effects of autonomy and surprise revealed by micro-phenomenological investigations. *Frontiers in psychology, 9*, 1704.

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., & Singer, D. (2009). *A mandate for playful learning in preschool: Applying the scientific evidence*. Oxford University Press.

Holmefur, M. M., & Krumlinde-Sundholm, L. (2016). Psychometric properties of a revised version of the Assisting Hand Assessment (Kids-AHA 5.0). *Developmental Medicine & Child Neurology, 58*(6), 618–624.

Khadka, J., Huang, J., Chen, H., Chen, C., Gao, R., Bao, F., Zhang, S., Wang, Q., & Pesudovs, K. (2016). Assessment of cataract surgery outcome using the Modified Catquest Short-Form Instrument in China. *PLoS ONE, 11*(10), 1–16. 10.1371/journal.pone.0164182.

Li, C. Y., Romero, S., Bonilha, H. S., Simpson, K. N., Simpson, A. N., Hong, I., & Velozo, C. A. (2016). Linking existing instruments to develop an activity of daily living item bank. *Evaluation & the Health Professions, 41*(1), 25–43. 10.1177/0163278716676873.

Linacre, M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. (Eds.) E. V. Smith,

& R. M. Smith (Eds.), Introduction to Rasch measurement: Theory, model and application *(pp. 258-278)*. JAM Press.

Linacre, J. M. (2021a). *Winsteps® (Version 4.8.0) [Computer Software]*. Beaverton, Oregon: Winsteps.com Retrieved January 1, 2021Available from https://www.winsteps.com/.

Linacre, J. M. (2021b). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com.

Magnuson, C. D., & Barnett, L. A. (2013). The playful advantage: How playfulness enhances coping with stress. *Leisure Sciences, 35*(2), 129–144.

Malec, J. F., Torsher, L. C., Dunn, W. F., Wiegmann, D. A., Arnold, J. J., Brown, D. A., & Phatak, V. (2007). The Mayo high performance teamwork scale: Reliability and validity for evaluating key crew resource management skills. *Simulation in Healthcare, 2*(1), 4–10.

Moeini, S., Rasmussen, J. V., Klausen, T. W., & Brorson, S. (2016). Rasch analysis of the Western Ontario Osteoarthritis of the Shoulder index - the Danish version. *Patient Related Outcome Measures, 7*, 173–181. 10.2147/PROM.S87048.

O'Connor, J. P., Penney, D., Alfrey, L., Phillipson, S., Phillipson, S. N., & Jeanes, R. (2016). The Development of the Stereotypical Attitudes in HPE Scale. *Australian Journal of Teacher Education, 41*(7), 70–87.

Proyer, R. T. (2017a). A new structural model for the study of adult playfulness: Assessment and exploration of an understudied individual differences variable. *Personality and Individual Differences, 108*, 113–122. 10.1016/j.paid.2016.12.011.

Proyer, R. T. (2017b). A multidisciplinary perspective on adult play and playfulness. *International Journal of Play, 6*(3), 241–243. 10.1080/21594937.2017.1384307.

Proyer, R. T., & Jehle, N. (2013). The basic components of adult playfulness and their relation with personality: The hierarchical factor structure of seventeen instruments. *Personality and Individual Differences, 55*(7), 811–816.

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

Shen, X. S., Chick, G., & Zinn, H. (2014). Playfulness in adulthood as a personality trait. *Journal of Leisure Research, 46*(1), 58–83. 10.1080/00222216.2014.11950313.

Singer, E. (2013). Play and playfulness, basic features of early childhood education. *European Early Childhood Education Research Journal, 21*(2), 172–184. 10.1080/1350293X.2013.789198.

Veas, A., Gilar, R., Castejón, J. L., & Miñano, P. (2016). Underachievement in Compulsory Secondary Education: a comparison of statistical methods for identification in Spain. *European Journal of Investigation in Health, Psychology and Education, 6*(3), 133–149.

Vogt, F., Hauser, B., Stebler, R., Rechsteiner, K., & Urech, C. (2018). Learning through play–pedagogy and learning outcomes in early childhood mathematics. *European Early Childhood Education Research Journal, 26*(4), 589–603.

Wong, M. H. Y., Fenwick, E., Aw. , A. T., Lamoureux, E. L., & Seah, L. L (2018). Development and validation of the Singapore thyroid eye disease quality of life questionnaire. *Translational Vision Science & Technology, 7*(5). 10.1167/tvst.7.5.14.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Yarnal, C., & Qian, X. (2011). Older-adult playfulness: An innovative construct and measurement for healthy aging research. *American Journal of Play, 4*(1), 52–79.