Jurnal Psikologi
Volume 47, Nomor 2, 2020: 104 – 120
DOI: 10.22146/jpsi.36860

ISSN 0215-8884 (Print)
ISSN 2460-867X (Online)
https://jurnal.ugm.ac.id/jpsi

The Use of the Partner Surveillance Scale in Instagram: Psychometric Evaluation Based on the Graded Response Model

Bambang Suryadi[1], *Muhammad Dwirifqi Kharisma Putra*[2]

[1,2]Faculty of Psychology, UIN Syarif Hidayatullah Jakarta
[2]Faculty of Psychology, Universitas Gadjah Mada

Submitted 10 July 2018 Accepted 06 February 2020 Published 24 August 2020

Abstract. The use of social media, especially Instagram, has become an increasingly powerful form of daily activity. This social media affects the romantic relationship of people, where people in relationships can conduct surveillance on the behaviors of their partner. This study provides an analysis of the psychometric properties of the Indonesian version of the Partner Surveillance Scale which contains 15 items and used a 4-point Likert scale format. The study recruited 214 female university students aged 17-23 years old, who used Instagram. The Graded Response Model (GRM) method was applied. As a result, the Indonesian version of the Partner Surveillance Scale was proved to have good psychometrics properties and had good fit to the GRM. All assumptions of GRM were met and the scale had high reliability. But, it should be noted that some items did not fit well with the model. The results of this study also provide an alternative to the use of Confirmatory Factor Analysis (CFA) in analyzing polytomous data with GRM. This study concluded that the psychometric properties of the Partner Surveillance Scale were good.

Keywords: graded response model; Instagram; partner surveillance scale

In the modern era, social media cannot be separated from everyday life. This phenomenon has attracted the attention of established psychological societies for example, the British Psychological Society (2012) and the American Psychological Association. The British Psychological Society (2012) published the paper "Guidance of the use of a social media by clinical psychologists". Three years later the APA published the article *"Social Media: A Contextual Framework to Guide Research and Practice"* (McFarland & Ployhart, 2015). The most updated research was conducted by Alhabash and Ma (2017) who found that the millennial generation choose social media because there is ease in communication with family and friends, or ease in seeking information and interacting with other people.

Some studies have shown that Instagram is one of the social media that was particularly appealing to millennials (Instagram, 2018; Pew Research Center, 2018; Rainie, Brenner, & Purcell, 2012). According to a survey conducted by Instagram (2018), the total number of Instagram users is 800 million with most users ageing between 18 and 24 years. Research from PEW Research Center (2018) showed that women tend to use social media more than men. Rainie et al.

Address for correspondence:
muhammad.dwirifqi@gmail.com

(2012) in their research, suggested that ease of uploading photos and videos on Instagram becomes one of the more attractive features of the platform among social media users.

In addition to the factors that lead to Instagram use among millennials, past research had shown that Instagram use is related to some psychological traits for example feelings of insecurity, fear of loss, popularity, happiness, and negative mood (Lup, Trub, & Rosenthal, 2015), intimate relationship satisfaction (Manvelyan, 2016), love, romanticism, jealousy (Ridgway & Clayton, 2016), hedonism (Casaló, Flavián, & Ibáñez-Sánchez, 2017), negative emotion (de Vries, Moller, Wieringa, Eigenraam, & Hamelink, 2017) as well as representations of women in hijab (Baulch & Pramiyanti, 2018). Other variables that were investigated by researchers include partner surveillance (Farrugia, 2013), partner-monitoring (Darvell, Walsh & Whire, 2011), social media surveillance (Brown, 2015) or interpersonal electronic surveillance (IES) which is often referred to as "Social media stalking" (Fox & Tokunaga, 2015).

According to Farrugia (2013), partner surveillance is the tendency for a person to monitor activities of the partner in social media, for example how the partner interacts with other people and how they comment on a user's post on Instagram. In relation to partner surveillance in social media, this behavior has the aim to identify whether a person is trustworthy or it may be simply due to jealousy (Muise, Christofides, & Desmarais, 2014). Furthermore, the motive may be to simply know the level of romanticism between partners (Serafinelli, 2017). However, monitoring is not always done effectively since users can adopt anonymous accounts, making it difficult for social media users to know who exactly is being monitored (Elphinston & Noller, 2011; Marshall, Bejanyan, Di Castro & Lee, 2013; Tokunaga, 2011).

Research in psychology had shown that partner surveillance constitutes the types of behaviors that are associated with trust and anxious attachment in intimate relations (Rodriguez, DiBello, Overup, & Neighbors, 2015) and is a development of the negative relational maintenance theory and the investment model (Tokunaga, 2016). The negative effects of partner surveillance is related to intimate partner violence (Rodriguez et al., 2015), depression and negative emotion (Marshall, 2012), as well as low quality relationships (Tokunaga, 2016). Research about partner surveillance continues to this day. One aspect which associates with its continuous growth is the availability of measurement instruments.

Results from the literature review had shown that there exist two measures that can be used to measure attitudes toward partner surveillance in social media namely: the Partner Surveillance Scale (PSS; Farrugia, 2013) as well as the Interpersonal Electronic Surveillance Scale (TIESS; Fox & Tokunaga, 2015). Although both of these measures have the same purpose, which is to observe behavior of social media users, there are some differences. TIESS cannot be used to measure aspects related with social media use apart from Facebook since the content of both measures contain features only found in Facebook and not in Instagram. However, PSS is quite flexible to adapt to the context of Instagram users. In addition, PSS uses a ranking format often known as the typical performance test. This form of measurement does not have a correct or incorrect answer because the goal is to

describe tendencies toward a specific trait (Hubley & Zumbo, 2013).

The use of Likert scales is aimed to measure psychological traits and it has become an important part in the use of advanced statistical methods, namely Item response theory or IRT (Adams, Wu & Wilson, 2012; Forero & Maydeu-Olivares, 2009). IRT consists of a compilation of statistical models that define the relationship between unobserved individual characteristics and item characteristics to predict a specific response towards an item in a scale (Baker & Kim, 2004). In applying the IRT, the focus of the analysis is on the item level and not on the overall scale. The focus on the item level makes it possible for the researcher to design, revise, and optimize the use of the scale to fit specific goals (de Ayala, 2009). Among the IRT models that can be used to analyze item scores in the form of a Likert scale, are Graded Response Model (GRM; Samejima, 1969), Rating Scale Model (RSM; Andrich, 1978) which is based on Rasch measurement, and factor analyses for example Item Factor Analysis (IFA; Wirth & Edwards, 2007). All three methods can be used to analyze item scores within an ordinal scale (for example Likert scale).

Although there exists research on the production of the Partner Surveillance Scale (Farrugia, 2013), there has not been any research in Indonesia to produce a similar scale with the characteristics, norms, and the local culture of Instagram users in Indonesia. Therefore this research was done to evaluate the psychometric properties of the Partner Surveillance Scale which contains 15 items and uses the 4 point Likert scale format. The current study was done among women using the Instagram social media with the GRM method.

Although GRM is nothing new and is available on numerous software, the application of psychological research in Indonesia is very limited when compared to application of CFM or RSM. This research is expected to give Indonesian researchers an introduction of the procedures to apply when interpreting the results of GRM analysis and analyzing the results of psychological traits. This research also produced an adapted scale which can be used for future research to test a range of other related variables.

Method

Participants

The research participants were students from Universitas Islam Negeri (UIN) Syarif Hidayatullah Jakarta. A total of 214 students were recruited with an age range of 17-23 years old and were selected based on non-probability sampling. This sampling technique was used due to time constraints of recruiting students who were active on their Instagram accounts. Other criteria for the eligibility of participation in this research include being female, actively using Instagram, first to fourth year students in the faculty of religion or general science, and were active students. Furthermore, all participation in this research was voluntary. Ethical clearance was obtained from the Institute of Research and Community Service (*Lembaga Penelitian dan Pengabdian Kepada Masyarakat*-LP2M), UIN Syarif Hidayatullah Jakarta.

Measuring instruments

The Partner Surveillance Scale (Farrugia, 2013) was developed at the Rochester Institute of Technology, United States of America. This instrument was developed

among a student sample and had good reliability (α = 0.84). This instrument was then translated to Bahasa Indonesia with the help of a professional translator using an online system. This instrument consisted of 15 items and used a Likert 4 point format with the following responses: Absolutely Disagree, Disagree, Agree, and Absolutely Agree, which is an adaptation of the original 5 point format. This adjustment was made based on suggestions from previous research to avoid disordered threshold (Adams, Wu & Wilson, 2012), especially for GRM models (García-Pérez, 2017). Disordered threshold occurs when the average response for the high category (e.g. absolutely agree) is lower compared to responses for the categories below it (e.g. absolutely agree). This would violate the assumption that higher agreement to a trait would correspond with a tendency to respond to the highest category in the scale (García-Pérez, 2017).

Data collection also considered demographic aspects of the respondents and included additional questions such as total followers in Instagram, age, income, relationship status: (a) in a relationship (b) was in a relationship but currently not in a relationship, and (c) had never been in a relationship. Four respondents who were not in a relationship were excluded from the analysis to minimize samples irrelevant to the context of the study. The items of the Partner Surveillance Scale can be seen in Appendix A. Following data collection, respondents with missing data were excluded from further analyses to avoid complex computing processes when dealing with missing data (for example modifying the estimation method).

Analysis procedure

Analysis was conducted using the Graded Response Model or GRM (Samejima, 1969). GRM is an IRT model that is used when the item scores are ordinal like Likert scales (Muraki, 1990; Samejima, 2016). Three assumptions must be met in GRM namely unidimensionality, local independence (Embretson & Reise, 2000) and monotonicity (de Ayala, 2009). This research tested three of those assumptions. Unidimensionality means that there is only one construct being measured, local independence refers to responses by the respondents that are statistically independent from responses to other items in a test, while monotonicity refers to the increase of a score corresponding to the higher level of the measured trait (de Ayala, 2009; Embretson & Reise, 2000).

In this research, the GRM model was estimated using marginal maximum likelihood (MML) using the program IRTPRO 3 (Cai, Thissen, & du Toit, 2015a). Basically the MML estimation method is used to estimate the standardized GRM model. The GRM equation is as follows:

$$P_{ik}^{*}(\theta) = \frac{exp[\alpha_i(\theta-\beta_{ik})]}{1+exp[\alpha_i(\theta-\beta_{ik})]} \quad (1)$$

In equation (1), P_{ik}^{*}, refers to the cumulative probability which is symbolized with *, α_i refers to the parameter of the discrimination power of the item i, β_{ik} refers to the parameter of threshold for option k on item i while θ refers to the estimation of the trait level of an individual. GRM is also known as *indirect IRT* because it is different from the dichotomous model, therefore the probability of choosing one response category cannot be conducted directly with formula (1) and so to calculate it for each response category formula (2) to (5) presented below can be applied

(Embretson & Reise, 2000; Samejima, 2016):

$$P_{i0}(\theta) = 1 - P_{i0}^{*}(\theta) \quad (2)$$

$$P_{i1}(\theta) = P_{i0}^{*}(\theta) - P_{i1}^{*}(\theta) \quad (3)$$

$$P_{i2}(\theta) = P_{i1}^{*}(\theta) - P_{i2}^{*}(\theta) \quad (4)$$

$$P_{i3}(\theta) = P_{i2}^{*}(\theta) - 0 \quad (5)$$

According to Standard 3.9 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), the overall model fit index or the item level fit must be reported when the IRT is used. The overall model fit indices used were M^2 and $RMSEA_2$. Should M^2 have $p > 0.05$ thus the data has unidimensional model fit (Maydeu-Olivares & Joe, 2006). If the $RMSEA_2$ has a value < 0.04, there is model fit for the data (Huggins-Manley & Han, 2017). Reliability in IRT uses the coefficient of marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984), which is analogous to the alpha coefficient in classical test theory (Reise, 1999). If the values are higher than 0.80, the instrument has good internal consistency (Petscher, Mitchell, & Foorman, 2015).

After testing fit for the overall model, we proceeded with testing model fit at the item level with the $S - \chi^2$ method. Items were said to have good fit when p from $S - \chi^2 > 0.05$ (Kang & Chen, 2008), while other authors suggest $p > 0.01$ (Stover, McLeod, Langer, Chen, & Reeve, 2019). If the item had low accuracy, the author would test the assumptions of local independence using the LD χ^2 method to identify the source of the item's problem.

The value of LD χ^2 which ranges from 3 to 5 shows that there is local dependence in the low category between one pair of items (Chen & Thissen, 1997), while the value > 10 shows that there is a major violation of the assumptions which might warrant modification of the overall model (Cai, Thissen, & du Toit, 2015b). If the whole model has good fit but there remains evidence of local dependence, the content of the items need to be re-evaluated and the author may consider dropping these items in future research (Depaoli, Tiemensma & Felt, 2018).

Results

The results of the analyses toward 15 items of the Partner Surveillance Scale (PSP) showed that the accuracy of the model fit was good (M^2 = 75.55, *df* = 60, p = 0.0848, $RMSEA_2$ = 0.03 and *marginal reliability* = 0.94). This finding showed that the assumption of unidimensionality was met (*p-value* from M^2 = 0.08 > 0.05; $RMSEA_2$ = 0.03 < 0.04). The reliability from PSS was 0.94 which indicates high internal consistency. Having established good fit with the data we proceed with an inter-pretation based on each item parameter (see Table 1).

Table 1 contains information on the discrimination power (slope), threshold, and index of item accuracy toward the model. All of the values of the threshold are ordered from the lowest threshold to the highest and this applies to all items. The patterns show that the assumption of *motononicity* had been met. None of the discrimination power of the items was negative which shows that the items were functioning well to differentiate people with a high trait and a low trait. We also found that there was a large slope for each item because they often ranged from 0.5 to 2.5. However, this was not a major issue since onlyvalue larger than 4.00 would indicate a serious problem (Edelen & Reeve, 2007), which was not the case in this research.

Table 1.

Estimation Results and Model Fit Index of Item Parameters

Item	Discrimination power (Slope)	Threshold			$S-\chi^2$		
		b1	b2	b3	χ^2	df	p-value
Item 1	2.94	-2.65	-1.21	0.60	37.78	23	0.0268*
Item 2	2.63	-2.58	-1.34	0.63	23.94	24	0.4665**
Item 3	3.32	-2.35	-1.13	0.46	35.94	23	0.0417*
Item 4	2.27	-2.69	-1.30	0.97	21.33	24	0.6201**
Item 5	3.12	-2.78	-0.94	0.56	32.53	22	0.0686**
Item 6	1.65	-3.29	-1.11	0.94	60.80	32	0.0016
Item 7	2.99	-2.81	-1.14	0.61	22.81	20	0.2976**
Item 8	3.00	-3.02	-0.99	0.32	35.05	22	0.0381*
Item 9	2.95	-2.31	-0.72	1.30	34.85	21	0.0292*
Item 10	2.41	-2.55	-0.94	0.93	33.73	27	0.1735**
Item 11	2.72	-2.27	-0.57	1.14	26.05	24	0.3495**
Item 12	2.63	-2.73	-0.94	0.74	36.04	24	0.0543**
Item 13	3.00	-3.36	-1.13	0.09	44.25	24	0.0071
Item 14	3.01	-1.72	-1.12	0.29	40.57	24	0.0185*
Item 15	2.93	-2.81	-1.14	0.38	26.82	22	0.2176**

Note: * *p-value* > 0.01, ** *p-value* > 0.05 shows item fit

Although there was good fit for the overall model with GRM, the information presented in Table 2 shows that some items have low accuracy based on the $S-\chi^2$ statistical test, namely items 6 and 13 which had $p < 0.05$ and $p < 0.01$. This means the items have low accuracy. However in contrast to CFA, items with low accuracy are not excluded from the analyses because this can have negative implications to the IRT model (Crişan, Tendeiro, & Meijer, 2017), indicating the different philosophy between GRM and models based on Rasch measurement (see, Andrich, 2004; Linacre, 2010). The results of the assumption test shows local dependence, which would be reported so that readers can understand why the items did not meet the criteria of good fit, to give information for future research (see Table 1).

Based on information from Table 2, there was a violation of the local independence assumption for two items which had low accuracy. Item 6 with items 11, 13, 14 and item 13 with items 3, 6, 9, 11, all these item pairs showed LD $\chi^2 > 3$, which exceeded the criteria of accuracy used. The violation of assumption between the items pairs are related to the unidimensionality assumption. However, violation of the local independence assumption in the research is not too strong as to violate the unidimensionality assumption. This is because the assumption on unidimensionality is most likely to be violated when the value of LD $\chi^2 > 10$ which would require a modification of the whole model (Cai, Thissen, & du Toit, 2015b).

Table 2.
LD χ^2 Statistic for Testing Assumption of Local Independence

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2	3.0													
3	*4.6*	*5.9*												
4	*3.1*	1.0	0.6											
5	1.9	1.9	1.2	-0.2										
6	2.7	1.6	1.4	0.5	*4.0*									
7	2.6	*4.4*	1.7	2.5	1.1	2.4								
8	0.6	1.3	1.8	0.2	1.5	1.6	1.6							
9	2.8	2.5	0.7	2.5	*3.8*	2.7	0.5	*5.1*						
10	2.4	1.1	-0.2	1.9	*3.3*	0.7	2.7	1.9	1.6					
11	2.8	0.6	0.6	-0.4	0.8	*5.1*	0.6	0.2	*4.5*	-0.7				
12	0.6	1.0	*3.5*	2.2	1.3	0.5	-0.1	*4.2*	0.6	1.2	*3.3*			
13	1.1	-0.1	*4.1*	3.0	1.6	*4.1*	0.7	1.3	*3.3*	0.3	*3.2*	0.8		
14	1.0	2.1	0.5	0.2	*3.8*	*4.5*	*4.2*	1.8	0.7	0.9	*4.0*	0.1	2.2	
15	1.6	1.8	1.1	0.9	0.3	1.1	0.3	2.1	-0.1	-0.3	1.4	0.7	1.4	1.6

The following information in figure 1 presents the ICC for each item and visually represents the characteristics of the items. Category 1 on all items, except for item 14 which is covered by categories that are in close proximity with it (option 0 and 2), function very well. For item 14, category 1 indicates that there is a continuum between disagree-absolutely agree. Participants tend to not choose these responses compared with other responses. We can see that the other categories are functioning well namely 0, 2, 3. This means that respondents with a low trait level tend to choose responses 0 and respondents with the high trait category tend to choose response category 3.

The use of GRM also produces estimates toward the total information curve (TIC). TIC supplies information related with estimation of the function of test information for each category of surveillance toward the partner. The x-axis shows the trait category of the respondent while the y-axis shows the overall information value and standard error (S.E). The value of the information is an inverse of the S.E., when S.E. is low than the information would be higher (See Figure 2).

As seen in figure 2, as far as the trait spans -3 logit to +1 logit, the magnitude of the information from the test was very high. This can be seen from S.E. which was below 0.30 meaning that this instrument is very informative on that category of partner surveillance. However, we must note that the test information curve was multimodal because it showed numerous peaks. The Information test curve that is multimodal needs specific attention from a statistical viewpoint for further interpretation. In general, the test information curve shows that PSS can give maximum information when it is used to measure respondents who exhibit low partner surveillance (-1 logit).

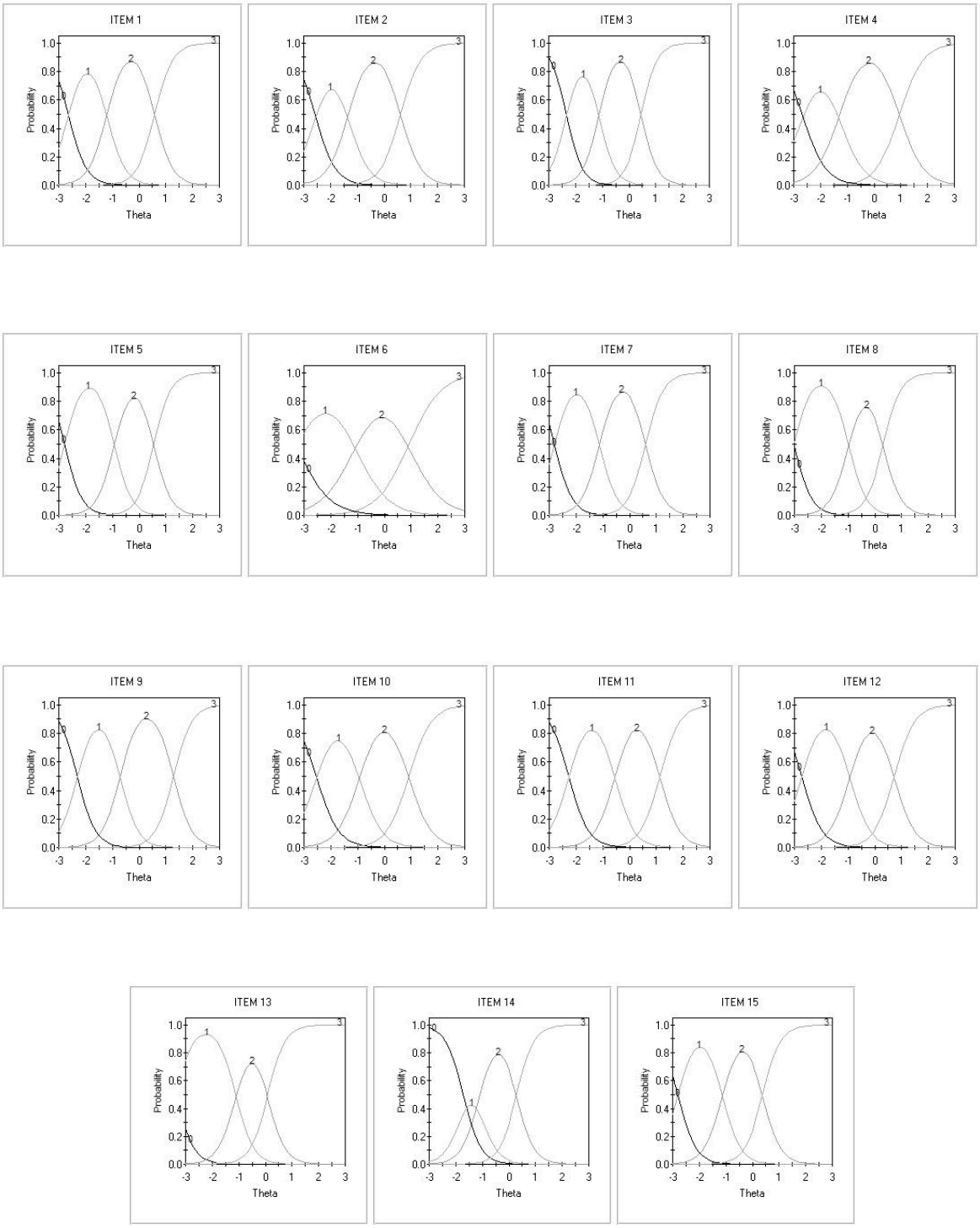

Figure 1. Curve of characteristics of 15 items of the Partner Surveillance Scale

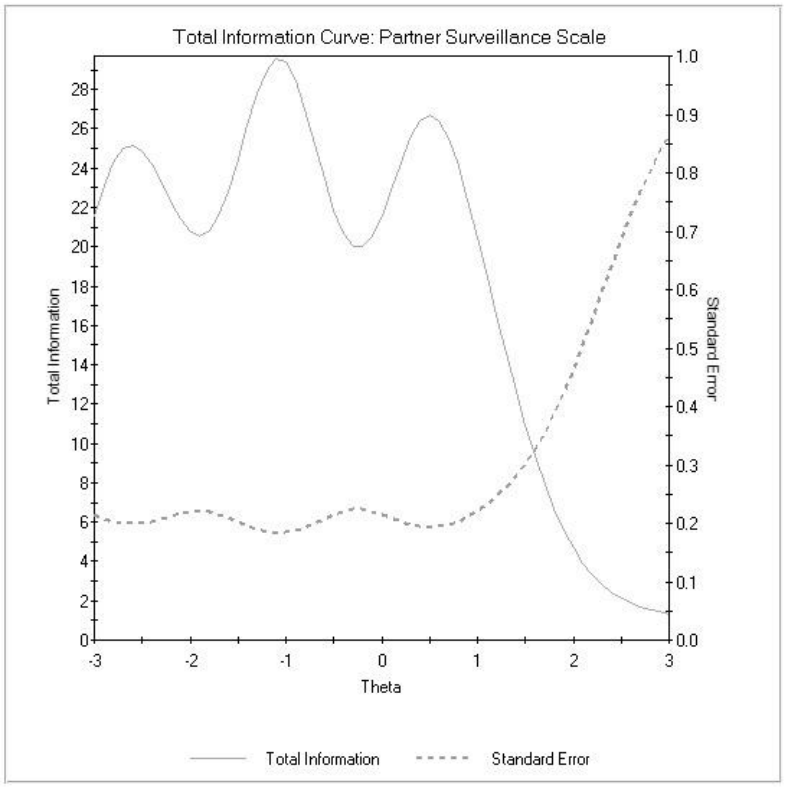

Figure 2. Curve of total information of 15 items of the Partner Surveillance Scale

Discussion

With the aim to evaluate the psychometric properties of the PSS, the results of the analyses using GRM showed that the PSS instrument has a unidimensional factor structure which means that the instrument measures one construct of partner surveillance on Instagram. If the model did not have good fit,the use of the GRM needs to be modified to other models for example multidimensional GRM (de Ayala, 1994) which accommodates different dimensions in the data.

The results of the GRM analysis showed that 2 of the 15 items of PSS have low accuracy toward the GRM model. These two items were further analyzed with regards to their violations of local independence to identify the core issue of the low accuracy. However, when the response categories were ordered from low to high, it showed that the assumption of monotonicity of PSS was met.

After further investigation we found that numerous item pairs had relationship values larger than the criteria namely item 6 with item 11, 13, 14, and item 13 with item 3, 6, 9, 11. The relationship that was observed may have been caused by the wording of the items (similar sentences used for the items) or it could be due to statistical error. In this research, the relationship that occurred was due to statistical artifacts since the relationship of the sentences between items did not have a pattern that we could conclude with certainty.

Although local dependence did not have a significant effect on the parameter estimates (Chen & Thissen, 1997), future research can further explore this issue with other statistical methods that can accommodate the existence of independent variables that could be accounted for (e.g.: relationship status and account ownership of the partner) as an example, by using IRT-C (Tay, Vermunt, & Wang, 2013).

Heterogeneity of the population has not been considered in this research and it was not controlled. In addition, the researcher could compare between models, for example compare the unidimensional GRM model with the multidimensional GRM model (de Ayala, 1994) or the Bifactor GRM (Cai, Yang, & Hansen, 2011) which accommodates the possibility of having more than one dimension or items that are part of more than one dimension. In this study, comparison toward numerous models have not been conducted even though this is advised by past research (Depaoli et al., 2018), and this is what becomes the limitation of the current research.

Furthermore, as a function of the response category, all response categories were ordered from low to high which means that the assumption of monotonicity was met and threshold disordering did not occur (see, Adams, Wu, & Wilson, 2012). However in item 14, the category Disagree had a low response when compared with the other three categories for example when observing the ICC of the items. This can be caused by the small number of respondents who answered to this category (García-Pérez, 2017). It was also found that the response category "Absolutely Disagree" for most of the items were located beyond the range of -3 of the trait category (very low), which is related to the low probability of this response category to be chosen by the respondents.

Based on further investigation toward the test information curve, we found that the curve was multimodal (had more than one peak). The function of the test information is very important in IRT, and if estimation toward the test information was inaccurate, there would be error in the interpretation of the test (Zhang, 2012). Although the research findings showed that overall there was large test information for both the low and high level partner surveillance, the multimodal curve showed that other factors were affecting test information that has not been accounted for in the current research. Some of these factors may include the size of the sample which may be too small to minimize bias on the parameter estimates of the items (Reeve, & Fayers, 2005; Zhang, 2012) as well as the high discrimination power of the item (Hambleton & Jones, 1994). The high item discrimination is in line with this research that showed that all items had high discrimination power, meaning that there is a high combination between discrimination power and threshold which creates variance of trait levels used in the computation of test information.

In numerous studies, it is common to find a relationship between test information function with reliability. However, caution must be made when using such an approach since the concept of reliability in IRT is different compared to classical approaches (Umar, 2012; 2014). Computation of reliability for GRM requires modification and specific computation (Samejima, 1994), and considering the complex computation required, this approach cannot be applied in the current study. However, reliability can be obtained through other methods, namely using marginal reliability with a value of 0.94 that describes a very good internal consistency.

This research also complements terms from past research namely "Social media stalker" (Fox & Tokunaga, 2015; Lyndon, Bonds-Raacke, & Cratty, 2011). This instrument can describe what is known as the "Instagram stalker" or " Instagram stalking", a term often used by the

younger generation in Indonesia. In addition, the researcher also contributes in understanding the index of model fit of IRT that is required in the report of analyses using IRT (Maydeu-Olivares, 2013, 2015). IRT based on statistical models like GRM have different philosophical leanings with Rasch-based models, for example RSM which operationalization is easier to apply (see, Andrich, 2004; Linacre, 2010). Both of them have their benefits and limitations. In addition to this comparison, this research shows that the application of the GRM model has assumptions that must be tested and this would be difficult for the researcher in the case of not having a good fit of the model or the items. The approach applied to overcome GRM models that did not have good fit is different for CFA, which can be managed by using the modification index (Sörbom, 1989). However, such things are not available in IRT models.

Conclusion

Based on the results of the research, it can be concluded that the Partner Surveillance Scale has good psychometric properties and can be used to measure partner surveillance. In addition, this measure also shows that it is unidimensional with a good model fit index and very good marginal reliability, although there were some violations of the assumption of local independence and low accuracy for two items toward the model. These issues need to be further explored in future research. Overall, the model can be applied in future research and can provide a technical description of what should be done for similar analyses methods. This research can become reference for researchers in psychology to conduct analyses using GRM methods.

Suggestion

Future research can conduct tests of the Partner Surveillance Scale on samples that are not only female, so we can obtain a different description of the functions of items across genders. Future research could also investigate the basic psychological variables related with partner surveillance behavior through Instagram which has not been covered in the current research.

Acknowledgments

The researcher extends gratitude to the editors and two anonymous reviewers that had given valuable feedbacks to this article.

Funding

The researcher did not receive funding in any form in the research or publication of this research.

Author's contribution

BS was responsible for the theoretical foundation of the paper, data collection and writing the manuscript. MDKP was responsible for the data analyses, drawing conclusion, and writing the article.

Conflict of interest

The authors declare there is not conflict of interest in this research.

Orcid id

Muhammad Dwirifqi Kharisma Putra. https://orcid.org/0000-0002-9383-7904

References

Adams, R. J., Wu, M. L. & Wilson, M. (2012). The Rasch rating model and disordered threshold controversy. *Educational and Psychological*

Measurement, 72(4), 547-573. doi: 10.1177/0013164411432166

AERA, APA, & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) Joint Committee on Standards for Educational and Psychological Testing (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alhabash, S., & Ma, M. (2017). A tale of four platforms: Motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students? *Social Media + Society*, 1-13, doi: 10.1177/2056305117691544

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573. doi: 10.1007/BF02293814

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigm? *Medical Care*, *42*(1), 1-16. doi: 10.1097/01.mlr.0000103528.48582.7c

Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press

Baulch, E., & Pramiyanti, A. (2018). Hijabers on Instagram: Using visual social media to construct the ideal Muslim woman. *Social Media + Society*, 1-15, doi: 10.1177/2056305118800308

Brown, I. (2015). Social media surveillance. In R. Mansell & P. H. Ang (Eds.), *The international encyclopedia of digital communication and society*, Hoboken, NJ: John Wiley & Sons, Inc.

Cai, L., Thissen, D., & du Toit, S. H. C. (2015a). IRTPRO for Windows (Version 3.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

Cai, L., Thissen, D., & du Toit, S. H. C. (2015b). *IRTPRO users guide.* Lincolnwood, IL: Scientific Software International

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis, *Psychological Methods, 16*(3), 221-248. doi: 10.1037/a0023350

Casaló, L. V., Flavián, C., & Ibáñez-Sánchez, S. (2017). Understanding consumer interaction on Instagram: The role of satisfaction, hedonism, and content characteristics. *Cyberpsychology, Behavior, and Social Networking, 20*(6), 369-375. doi: 10.1089/cyber.2016.0360

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational Behavioral Statistics, 22*(3), 265–289. doi: 10.3102/10769986022003265

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement*, *41*(6), 439-455. doi: 10.1177/0146621617695522

Darvell, J., Walsh S. P., & White, K. M. (2011). Facebook tells me so: Applying the theory of planned behavior to understand partner-monitoring behavior on Facebook. *Cyberpsychology, Behavior and Social Networking, 14*(12), 717-722. doi: 10.1089/cyber.2011.0035

de Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*(2),

155-170. doi: 10.1177/014662169401800205

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

de Vries, D. A., Moller, A. M., Wieringa, M. S., Eigenraam, A. W. & Hamelink, K. (2017). Social comparison as the thief of joy: Emotional consequences of viewing strangers' Instagram posts. *Media Psychology, 22*(2), 222-245. doi: 10.1080/15213269.2016.1267647

Depaoli, S., Tiemensma, J. & Felt, J. M. (2018). Assessment of health surveys: fitting a multidimensional graded response model. *Psychology, Health & Medicine, 23*(1), 13-31. doi: 10.1080/13548506.2018.1447136

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modelling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16(1)*, 5–18. doi: 10.1007/s11136-007-9198-0

Elphinston, R. A. & Noller, P. (2011). Time to face it! Facebook intrusion and the implications for romantic jealousy and relationship satisfaction. *Cyberpsychology, Behavior and Social Networking, 14*(11), 631-635. doi: 10.1089/cyber.2010.0318

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

Farrugia, R. C., (2013). *Facebook and relationships: A study of how social media use is affecting long-term relationships* (Unpublished Master's Thesis). Rochester, NY: Rochester Institute of Technology.

Forero, C., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275–299. doi: 10.1037/a0015825

Fox, J. & Tokunaga, R. S. (2015). Romantic partner monitoring after breakups: Attachment, dependence, distress, and post-dissolution online surveillance via social networking sites. *Cyberpsychology, behavior, and social networking, 18*(9), 491-498. doi: 10.1089/cyber.2015.0123

García-Pérez, M. A. (2017). An analysis of (dis)ordered categories, thresholds, and crossings in difference and divide-by-total IRT models for ordered responses. *The Spanish Journal of Psychology, 20*(10), 1-27. doi: 10.1017/sjp.2017.11

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360. doi: 10.1111/j.1745-3984.1984.tb01039.x

Hambleton, R. K., & Jones, R. W. (1994) Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education*, 7(3), 171-186. doi: 10.1207/s15324818ame0703_1

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (pp. 319). Washington, D.C.: American Psychological Association Press.

Huggins-Manley, A. C. & Han, H. (2017). Assessing the sensitivity of weighted least squares model fit indexes to local dependence in item

response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 331-340. doi: 10.1080/10705511.2016.1247355

Instagram. (2018). *Instagram statistics.* Retrieved from instagram.com/press (8 June 2018)

Kang, T., & Chen, T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Education Measurement, 45*(4), 391–406. doi: 10.1111/j.1745-3984.2008.00071.x

Linacre, J. (2010). Two perspectives on the application of Rasch models. *European Journal of Physical and Rehabilitation Medicine*, *46*(2), 309-310.

Lup, K., Trub, L., & Rosenthal, L. (2015). Instagram #instasad? Exploring associations among Instagram use, depressive symptoms, negative social comparison, and strangers followed. *Cyberpsychology, Behavior, and Social Networking, 18*(5), 247–252. doi: 10.1089/cyber.2014.0560

Lyndon, A., Bonds-Raacke, J., & Cratty, A. D. (2011). College students' Facebook stalking of ex-partners. *Cyberpsychology, Behavior, and Social Networking, 14*(12), 711-716. doi: 10.1089/cyber.2010.0588

Manvelyan, C. (2016). Pics or it didn't happen: Relationship satisfaction and its effects on Instagram use. *Colloquy*, *12*, 87-100.

Marshall, T. C. (2012). Facebook surveillance of former romantic partners: associations with post breakup recovery and personal growth. *Cyberpsychology, Behavior, and Social Networking, 15*(10), 521–526. doi: 10.1089/cyber.2012.0125

Marshall, T. C., Bejanyan, K., Di Castro, G. & Lee, R. A. (2013). Attachment styles as predictors of Facebook-related jealousy and surveillance in romantic relationships. *Social Psychology*, *20*(1), 1-22. doi: 10.1111/j.1475-6811.2011.01393.x

Maydeu-Olivares, A. (2013). Why should we assess the goodness-of-fit of IRT models? *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 127-137.

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111–127). New York: Routledge.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732. doi: 10.1007/s11336-005-1295-9

McFarland, L. A. & Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of Applied Psychology, 100*(6), 1653-1677. doi: 10.1037/a0039244

Muise, A., Christofides, E., & Desmarais, S. (2014). "Creeping" or just information seeking? Gender differences in partner monitoring in response to jealousy on Facebook. *Personal Relationships, 21*(1), 35-50. doi: 10.1111/pere.12014

Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement, 14*(1), 59-71. doi: 10.1177/014662169001400106

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: an illustration of conditional item

response theory using a computer-administered measure of vocabulary. *Reading and Writing*, *28*(1), 31-56. doi: 10.1007/s11145-014-9518-z

Pew Research Center. (2018). *Social media use in 2018.* Washington, DC: Pew Research Center

Rainie, L., Brenner, J., & Purcell, K. (2012). *Photos and videos as social currency online.* Retrieved from www.pewinternet.org/2012/09/13/photos-and-videos-as-social-currencyonline/ (8 June 2018)

Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In P. M. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 55-73), Oxford, UK: Oxford University Press

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219– 241). Mahwah, NJ: Erlbaum.

Ridgway, J. L., & Clayton, R. B. (2016). Instagram unfiltered: Exploring associations of body image satisfaction, Instagram #Selfie posting, and negative romantic relationship outcomes. *Cyberpsychology, Behavior and Social Networking, 19*(1), 2-7. doi: 10.1089/cyber.2015.0433

Rodriguez, L. M., DiBello, A. M., Overup, C. S., Neighbors, C. (2015). The price of distrust: trust, anxious attachment, jealousy, and partner abuse. *Partner Abuse*, *6*(3), 298-319. doi: 10.1891/1946-6560.6.3.298

Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores,* Psychometrika Monograph, 17. Richmond, VA: Psychometric Corporation

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229-244. doi: 10.1177/014662169401800304

Samejima, F. (2016). Graded Response model. In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 85-100), Berlin: Springer

Serafinelli, E. (2017). Analysis of photo sharing and visual social relationships: Instagram as a case study. *Photographies*, *10*(1), 91-111. doi: 10.1080/17540763.2016.1258657

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384. doi: 10.1007/BF02294623

Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W-H., & Reeve, B. B. (2019). State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *Journal of Patient-Reported Outcomes*, *I*(1), 50. doi: 10.1186/s41687-019-0130-5

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the Item Response Theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing, 13*(3), 201-222. doi: 10.1080/15305058.2012.692415

The British Psychological Society. (2012). *e-Professionalism: Guidance on the use of social media by clinical psychologists.* Leicester, UK: The British Psychological Society

Tokunaga, R. S. (2011). Social networking site or social surveillance site? Understanding the use of interpersonal electronic surveillance in romantic relationships. *Computers in Human Behavior*, *27*(2), 705-713. doi: 10.1016/j.chb.2010.08.014

Tokunaga, R. S. (2016). Interpersonal surveillance over social network sites: applying a theory of negative relational maintenance and the investment model. *Journal of Social and Personal Relationships*, *33*(2), 171-190. doi: 10.1177/0265407514568749

Umar, J. (2012). Mengenal lebih dekat konsep reliabilitas skor tes. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia* (JP3I), *2*(2), 126-140.

Umar, J. (2014). Kerancuan dalam penggunaan istilah "construct reliability". *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia* (JP3I), *3*(4), 393-400.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58-79. doi: 10.1037/1082-989X.12.1.58.

Zhang, J. (2012). The impact of variability of item parameter estimators on test information function. *Journal of Educational and Behavioral Statistics, 37*(6), 737-757. doi: 10.3102/1076998612458321

Appendix A

Partner Surveillance Scale Indonesian Version (Back translated to English)

No.	Item
1	I trust my partner
2	I trust my partner's online activities *
3	I investigate my partner's Instagram profile
4	I investigate my partner's Instagram profile to monitor her/his online activities *
5	I investigate my partner's Instagram profile to see her friend's online activities *
6	I am sure my partner follows the Instagram account of his/ her ex-boy/girlfriend
7	I am irritated when knowing that my partner follows the Instagram account of her/his ex-boy/girlfriend *
8	I am happy when I see my boy/girlfriend uploads a photo/video about me and him and tags my account
9	I am happy when my partner uploads a photo/video about me and him on his account
10	I am happy when I upload a photo which shows me together with him *
11	I am happy when my partner uploads a photo of us together *
12	I know everyone who checks my partners Instagram account *
13	I am happy when I see other people uploading content showing their relationship on Instagram
14	I am happy when I see photos uploaded on Instagram by people who are in relationships
15	I personally feel, that people in relationships should show their happiness online through their Instagram account *

* = Items with high accuracy

Each item in PSS consists of 4 answer options: AD (Absolutely Disagree) = 1, D (Disagree) = 2, A (Agree) = 3, AA (Absolutely Agree) = 4. PSS has a total score range of 15 to 60. Higher score on PSS indicates a person's agreement toward performing partner surveillance behaviors.