

Final Validity Manual
(PSYC 746 McGill University)

Gyeongcheol Cho

I. Introduction and Background

Van Holten, a company that mainly supplies pickled products to the market, decided to rebrand their pickled products to enter the global market. According to the pilot study they conducted, however, their flagship product, named “Big Papa”, had several issues: for instance, customers had difficulty finding the product’s key information on its cover, and the image on the cover turned out to be unpleasant to some customers. Thus, the company has hired market researchers at EvilCorp to resolve these issues.

To have a better understanding of the market for pickled products and develop a successful marketing strategy, EvilCorp thought that it is crucial to identify people having pickle fanaticism around the globe. For the identification, they had to develop a precise measure of pickle fanaticism (PF) in advance, because an instrument for measuring pickle fanaticism did not exist at that time. Once the measure was developed and validated, they planned to find pickle fanatics (i.e., people having a high level of pickle fanaticism) by using the measure, to investigate their responses to the pickled products Van Holten newly launched, and to utilize those people as social media influencers of Van Holten’s pickled products in American, European, and Asian markets.

The literature they reviewed said that pickle fanaticism refers to “the general zeal for pickled products”, which is potentially comprised of three sub-domains: (1) “a strong desire to eat pickled products”, (2) “the extreme liking of pickled products”, and (3) “the general feeling of needing to evangelize to others about the benefits of pickled products” (see Assignment 1, 2021). Thus, EvilCorp made 33 candidate items of pickle fanaticism, each of which is expected to measure one of the domains of pickle fanaticism. EvilCorp entrusted me to assess the validity of these items, so I conducted a validation study according to the guidelines proposed in the psychometrics literature.

II. Qualitative Item Review

I conducted the qualitative item review for the 33 items of pickle fanaticism. This review mainly aimed to investigate whether each item's content was indicative of its target domain and what cognitive process (e.g., interpreting the item and making/adjusting their response) was involved when people responded to the item. The former can provide validity evidence based on test content, whereas the latter can yield validity evidence based on cognitive response.

To obtain validity evidence based on test content, I, as a domain expert, reviewed the item contents. As the test items are measured on the Likert scale, I examined the item contents according to the Edwards's (1957) guideline on the Likert scale. The contents of each item can be found in Table 1. The results revealed that many items were inconsistent the Edwards's (1957) guideline. Specifically, candidate items 3, 4, 7, 12, 23, 26, and 31 (e.g., "Pickles are made with vinegar") were not relevant to their target domains. Some candidate items (2, 9, 13, 14, 17) were too extreme or contained universals/leading words so that almost all the respondents were (not) likely to endorse those items (e.g., "I only eat pickles"). Also, many items (15, 18, 26, 31) were found to be negatively phrased (e.g., "I don't mind pickles"), which would make respondents difficult to disagree with those items. Items 3, 7, and 9 contained difficult vocabulary (e.g., "delectable sustenance"), so that some people could not understand the meaning of the items. The statement of item 4 ("I dream about pickles") seemed to be ambiguous; both respondents extremely liking pickled products and those hating pickled products might endorse this item. Item 2 asked respondents' experience in the past (i.e., 30 years ago), to which young people could not provide a proper answer. Lastly, item 24 contained multiple conditions, having respondents difficult to select their answers when they disagree with some of the conditions.

To obtain validity evidence based on response process, I conducted Think-aloud interviews with two people. Both were male engineers in their early thirties, who used English as their second language.

I had two prior themes for items that might confuse respondents: (1) multiple interpretations and (2) difficult vocabulary. The interviewees' description confirmed that several items had such issues, but the first issue occurred from the items that I had not expected: items 19, 22, and 29 (e.g., "Pickles are a unique food"). Interviewees thought that these items asked about whether the statements are true or not in general, rather than about their own preference. Both respondents had difficulty understanding the meaning of items 3, 7, and 9, as expected.

Once the Think-aloud interview was complete, I conducted retrospective probes to further grasp the interviewees' response process. I asked them a couple of additional questions based on their answers in the interview. For example, I asked an interviewee why he chose "disagree" rather than "strongly disagree" for item 25 ("My family should eat pickles regularly"), even though he argued that he couldn't find any positive side from pickles. By this procedure, I could identify one important additional theme: unclear meaning of the Likert scale. The current Likert scale itself turned out to be very confusing to those who don't mind pickles at all. For example, even if both interviewees didn't like pickles at all, one interviewee chose "Neither agree nor disagree" for item 22 ("I like pickles a lot") but the other chose "Strongly disagree" for the item. The first interviewee thought that the choice of "Strongly disagree" indicated he disliked pickles, whereas the second interviewee interpreted that such choice simply indicated he disagreed with the item statement itself. A similar issue occurred to the responses to item 25. One interviewee had difficulty responding to item 25 because disagreeing with the item could be interpreted not just that he simply disagreed with the item, but also that he thought his family should 'not' eat pickles regularly.

The review of item contents showed that many test items might not be appropriate for measuring the domains of pickle fanaticism. The Think-aloud interviews and retrospective probes confirmed it in terms of respondents' cognitive process and additionally suggested that the Likert scale used for pickle

fanaticism could be ambiguous. Considering the results of the qualitative item review, I proposed to remove or revise some items as shown in Table 1. Also, I proposed to clarify the meaning of “Disagree” option on the Likert scale with an illustration in the questionnaire; for instance, I would add a description like “strongly disagreeing with item 22 (“I like pickles a lot”) should be interpreted that the respondent hates pickles a lot.”

Table 1. Recommendations for the 33 candidate items of the Pickle Fanaticism questionnaire

Domain	Statement	Decision	Suggested revision
1	1. I think about eating pickles at most meals	keep	-
	2. Over the past 30 years I have eaten thousands of pickles	remove	-
	3. I often contemplate the role of pickles in a post-modern society	remove	-
	4. I dream about pickles	remove	-
	5. I prefer to eat other snacks over pickles	keep	-
	6. My go to snack is a pickle	keep	-
	7. I would rather eat dill pickles than sweet pickles	remove	-
	8. I eat pickles often	keep	-
	9. Pickles are delectable sustenance	remove	-
	10. I would like to eat pickles everyday	revise	I would like to eat pickles regularly
	11. I avoid eating pickles	keep	-
2	12. Pickles are made with vinegar	remove	-
	13. I only eat pickles	remove	-
	14. I always eat pickles for breakfast	remove	-
	15. I do not like pickles much	revise	Pickles are far from my type
	16. I like pickles less than other foods	revise	I like pickles less than other side dishes
	17. Pickles taste extremely good	revise	Pickles taste good to me
	18. I don't mind pickles	keep	-
	19. Pickles are a unique food	remove	-
	20. Pickles taste bad	revise	For me, the taste of pickles is bad
	21. Pickles are delicious	keep	Pickles are delicious to me
	22. I like pickles a lot	keep	-
3	23. I have many friends who like pickles	remove	-
	24. Weekly I make sure to tell a friend about pickles, the different kinds of pickles, where you can buy them, and how much they cost	revise	I tell my friends about the benefits of pickles at times.
	25. My family should eat pickles regularly	revise	I recommend my family to have pickles
	26. I don't like my friends who don't eat pickles	remove	-
	27. Everyone should eat pickles	revise	Some people need to eat pickles
	28. Pickles are great gifts	remove	-
	29. I like to tell people about my favorite pickles	keep	-
	30. I want to share my love of pickles with the world	keep	-
	31. My friends should not eat pickles	revise	I want my friend to avoid eating pickles
	32. I'm secretive about my pickle habits	remove	-
	33. I recommend pickles to people I know	keep	-

* Domain 1 = ‘Strong desire to eat pickled products’, Domain 2 = extreme liking of pickled products, and Domain 3 = feeling the need to evangelize about the benefits of pickles.

III. Descriptive Item Analysis

I conducted the descriptive item analysis for the 33 items of pickle fanaticism. In this analysis, I mainly investigated the overall response pattern for each item (e.g., sparsity, skewness), its sample statistics (e.g., mean, standard deviations, minimum and maximum values), and correlations among the items having the same target domain. Through this procedure, I could obtain three types of validity evidence: evidence based on content validity, evidence based on cognitive response process, and evidence based on the internal structure.

Specifically, I analyzed data collected from 563 respondents. The analysis began with drawing histograms for each item and obtaining their descriptive statistics. At this stage, I could empirically check whether people responded abnormally to some items, whether people partially used the scale, and whether the distributions of item scores were skewed. Then, I interpreted each result and sought to find validity evidence based on content validity and cognitive process. For example, if the distribution of scores of an item was far from the prior expectation, it may imply that respondents could have interpreted the item content in a different way from what was intended, or the response to the item would have been affected by other external factors that were not relevant to the target domain. Also, if some categories of an item had no response or the distribution of scores of the item was highly skewed, it may imply that the item could have covered an extremely narrow range of levels of the target construct. Next, I derived the correlation matrix of items to obtain the validity evidence based on internal structure. As the items were expected to measure the same domain of a construct, they had to be highly correlated. Lastly, I examined the item-total correlations for each item. The low item-total correlation of an item indicated the possibility that the item may not measure the same construct as the others.

By the analysis, I could identify the items supported by validity evidence along with problematic items for each domain. The results and recommendations are summarized in the following table.

Table 2. Results from descriptive item analysis and recommendations.

Domain	Item number	Pattern	Interpretation/Implications	Recommendation
1,2	13, 14	These items had the sparsest responses in some categories.	These items may cover a narrow range of construct levels, due to the universals in their statements.	Remove the universals in the items
	15	This item was positively skewed, even though all the other negatively worded items were negatively skewed.	Only item 15 was negatively phrased. Other negatively worded items were negatively oriented. The negative phrase in item 15 may confuse respondents.	Change "I do not like pickles much" to "I hate pickles".
	10,18,9	Their scores were evenly distributed with high SD, compared to other items' scores.	These items may not measure the same construct as others or be affected by other factors additionally.	Remove the items
	12,19	These items were weakly correlated with other items	These items may not measure the same construct. Item 12 is factual, and item 19 can be interpreted as factual.	Remove the items
	5,11,15, 16,20	These items were negatively correlated with the item total scores.	These items were negatively worded, lowering the correlation with the item total scores.	Reverse their scores.
	8,10,17, 21,22	These items were highly correlated with each other	These items would measure the same underlying construct.	Keep these items and use one of the reference items.
3	24,26	These items were extremely positively skewed.	This would be because the statements in the two items are too extreme so that few people would have endorsed the statements.	Remove the items
	23,32	These items were weakly correlated to the other items.	The two items may not measure the same construct as the others.	Remove the items
	28,29, 30,33	These items were strongly correlated with each other	These four items would measure the same construct. Considering their contents, these items seem to be a good indicator of Domain 3	Keep these items and use one of the reference items.
	31	This item was negatively worded. Recoding this item made the item positively correlated with others	The item would measure the same construct as others	Recode the item

	32	The item was negatively worded. However, recoding this item did not make the item positively correlated with others	The item would not measure the same construct as others	Remove the items
	23,31,32	The items have the lowest item-total correlations. Removing those items substantially increased the average inter-item correlations.	These items may not measure the same construct as others	Consider removing some of the items.

* Domain 1 = 'Strong desire to eat pickled products', Domain 2 = extreme liking of pickled products, and Domain 3 = feeling the need to evangelize about the benefits of pickles.

In Summary, the descriptive item analysis revealed that many items lacked validity evidence.

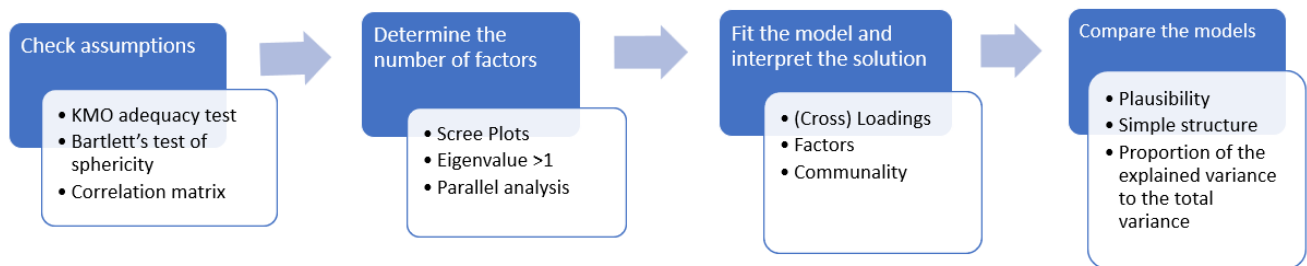
Thus, I proposed to exclude items 9,10,12,18,19, 23, 24, 26, and 32 from the item pool for pickle fanaticism. Also, I recommended recoding the negatively worded items for using the total item scores as measurements of pickle fanaticism.

IV. Exploratory Factor Analysis

Exploratory factor analysis (EFA) enables researchers to study how many factors underlie the items and to explore the most plausible measurement model for data without relying on prior knowledge. In general, researchers expect each item to load on only one factor or their measurement model to achieve the simple structure. Accordingly, EFA seeks to find the solution satisfying the simple structure by using the so-called rotation algorithm. Once such a solution is found and the solution is theoretically plausible, researchers use the results as validity evidence based on internal structure.

I conducted the exploratory factor analysis (EFA) for the 11 items of pickle fanaticism. The general procedure of applying EFA is displayed in the following figure.

Figure 1. The general procedure of applying EFA



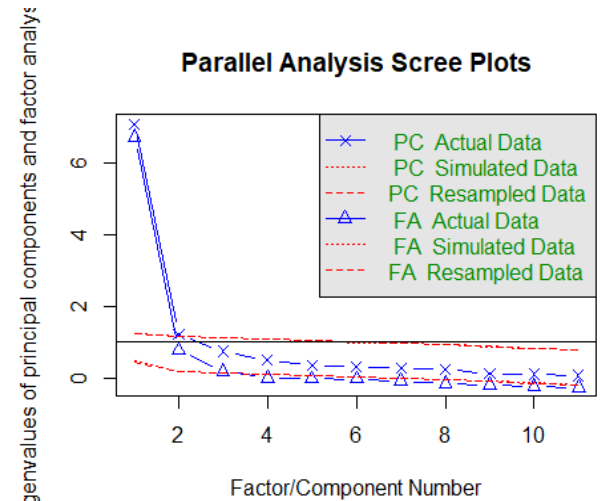
I initially conducted a KMO adequacy test and computed the MSA that evaluates the amount of shared variance of indicators. If MSA is close to 1, it means that items share the variance sufficiently. Overall MSA was .93, indicating the marvelous MSA value according to Kaiser's rules of thumb (Kaiser, 1960). On the other hand, Bartlett's test of sphericity (Bartlett, 1951) quantifies the discrepancy between the sample correlation and identity matrices and statistically tests whether the discrepancy is statistically significant. Bartlett's test of sphericity showed that there was a significant difference between the sample correlation and identity matrices ($\chi^2(55) = 5936.865$, $p = .000$), indicating that the sample correlation matrix was not equivalent to the identity matrix. Considering the correlation matrix of the

items, I expected the items to load on two factors. Two sets of items (11,17,20,21, and 22 for factor 1; 28,29, and 30 for factor 2) showed relatively high correlations with each other within each set.

To decide the number of factors to extract, I checked the scree plot, the eigenvalues, and the results of the parallel analysis. The scree plot went off into a flat tail when the number of factors was three, suggesting two factors. Only one factor had the eigenvalue greater than 1, which means that one factor was recommended. In the parallel analysis, factor number 3 was the largest number whose eigenvalue was higher than that of resampled/simulated data, indicating that the three-factor

solution was recommended. As the sample size was not small and the PA tends to suggest more factors than the actual number with larger sample sizes (Warne & Larsen, 2014), I assumed 3 factors as the maximum number of factors in the EFA model. Taken together, I had to examine the EFA models from 1 to 3 factors.

In the analysis, I examined two-factor and three-factor EFA models and compared their results. I adopted Oblimin rotation, in which factors can be correlated because domains of pickle fanaticism were likely to be associated. In the three-factor model, items 11, 17, 20, 21, 22, and 33 loaded on factor 1, items 28, 29, 30, and 33 loaded on factor 2, and items 25 and 27 loaded on factor 3. Item loadings for each factor were larger than .32, indicating that 10% of the item variance was explained by the factor. However, factor 3 was not viable because only two items loaded on the factor, though at least three items per factor are generally recommended. Furthermore, it seemed that factors 2 and 3 were simply different in the severity of the construct while commonly corresponding to the desire to evangelize about



the benefits of pickled products. The proportion of the explained variance of the items to the total variance was .75.

On the other hand, in the two-factor model, items 11, 17, 20, 21, 22, and 25 loaded on factor 1, whereas items 25, 27, 28, 29, 30, and 33 loaded on factor 2. Both factors were viable since both had more than 3 indicators. The first factor could be interpreted as ‘pickle liking’, whereas the second factor as ‘pickle evangelism’. The proportion of the explained variance of the items to the total variance was .70. The correlation between the two factors was .65, indicating that there was a strong, positive association between ‘pickle liking’ and ‘pickle evangelism’

Between the two solutions, I recommended selecting the two-factor solution. As mentioned above, factors 2 and 3 in the three-factor model were essentially identical and factor 3 was not viable because only two items loaded on the factor. Also, the difference in the proportion of the explained variance of the items to the total variance was not substantial (.05).

V. Reliability Analysis

Reliability refers to the “consistency of measurements across conditions.” (Bandalos, p.157).

Researchers should choose the proper measure of reliability depending on the consistency of their interests. The four types of reliability measure are summarized in Table 3.

Table 3. Four types of reliability measure

Type of reliability	Consistency of interest
Test-retest reliability (or coefficients of stability)	Consistency of scores over time
Alternative forms reliability (coefficients of equivalence)	Consistency of scores across test forms
Internal consistency (e.g., coefficient's alpha)	Consistency of scores across items
Measures of interrater agreement	Consistency across raters

In this validation study, I had to examine how consistent scores were across items for each domain. This consistency is called *internal consistency*, which is a type of validity evidence based on internal structure. The high level of internal consistency reliability can serve as validity evidence under the condition that systematic error does not exist. In this condition, the reliability becomes equivalent to the proportion of the true score variance to the observed score variance.

As a measure of internal consistency reliability, one can use split-half reliability and Cronbach's alpha. However, I used the latter only, because the former has two disadvantages that (1) only half of the items are used to calculate internal consistency and (2) the choice on how to split the items alters the estimate. It is proven that Cronbach's alpha is equivalent to the average of the all-possible split-half reliability estimates. Based on the EFA results obtained in the previous section, I regarded a set of items 11, 17, 20, 21, 22, and 25 as a measure of factor 1 ('pickle liking') and a set of items 27, 28, 29, 30, and 33 as a measure of factor 2 ('pickle evangelism'). The α estimates for each subscale were as follows.

For factor 1, $\alpha = .95$ (95% CI = [.95, .96])

For factor 2, $\alpha = .86$ (95% CI = [.84, .88])

The α estimate for the first subscale was greater than .90 and its 95% CI also did not include .90. It indicates that the first subscale would be considered acceptable regardless of its purposes. On the other hand, the α estimate for the second subscale and its 95% CI were between .80 and .90. It implies that this subscale would be considered acceptable for research purposes, not for clinical purposes. The reason why the first subscale had a higher α level than the second subscale would be that the first subscale had a larger number of items (6) than the second one (5) and the items for the first subscale had a higher inter-item correlation on average ($r=0.769$) than those for the second subscale ($r=0.554$). I examined whether the reliability of the second subscale could be improved by dropping some of its items. Removing item 27 turned out to increase the α level from .861 to .863 but the difference was negligible.

To demonstrate the usage of the finalized subscales with their α levels, I searched for a group of people whose scores on the pickle evangelism subscale was one standard deviation (SD) beyond the mean and regarded the group as having a high level of pickle evangelism. As the pickle evangelism subscale had a mean of 13.03 and an SD of 5.45 so that people whose scores on the pickle evangelism subscale were more than 18.483 belonged to this group. The cutoff score (18.483) could be justified in that people having the lowest pickle evangelism score in this group had the 95% CI of [14.50, 22.47], which did not include the subscale mean (13.03). The 95% CI of the item total score can be obtained by “the mean $\pm 1.95 \times \text{SEM}$ (standard error of measurement)” and the SEM is equivalent to the SD $\times \sqrt{1 - \alpha}$.

In summary, the two subscales of Pickle fanaticism had an α level greater than .80. Considering the purpose of the development of this scale (not clinical purpose), this reliability level is sufficiently high and can be used as a piece of validity evidence based on internal structure. However, users need to

be cautious of interpreting α in terms of validity evidence because unknown systematic errors might increase the α level of the subscales substantially while making the subscales less valid.

VI. Correlation Analysis

The purpose of this correlation analysis is to investigate validity evidence based on relations to other variables. If a construct is properly measured, its correlations with other variables including observed variables and constructs will be consistent with the theoretical expectation, under the assumption that the theory is valid. According to which type of variables or how large correlation is expected, validity evidence based on relations on other variables can be classified as follows.

Table 4. Four types of validity evidence based on relations to other variables

	Relevant variables	Expected correlation
Predictive	Observed variables measured in the future	medium to large
Concurrent	Observed variables measured at the same time point	medium to large
Convergent	Constructs	medium to large, but not too large
Discriminant	Constructs	zero to small

For this analysis, EvilCorp finally refined/selected 9 items for the final Pickle Fanaticism Scale (11,17,20,21 for ‘Extreme liking of pickled products’ or likingScore; 25, 27, 28, 29, 30, 33 for ‘Feels the need to evangelize about the benefits of pickles’ or evangScore), collected data again with other relevant variables, and asked me to examine the validity evidence based on the relations with those variables. The list of the relevant variables, my hypotheses, the validity evidence I could obtain from each of them is summarized as follows.

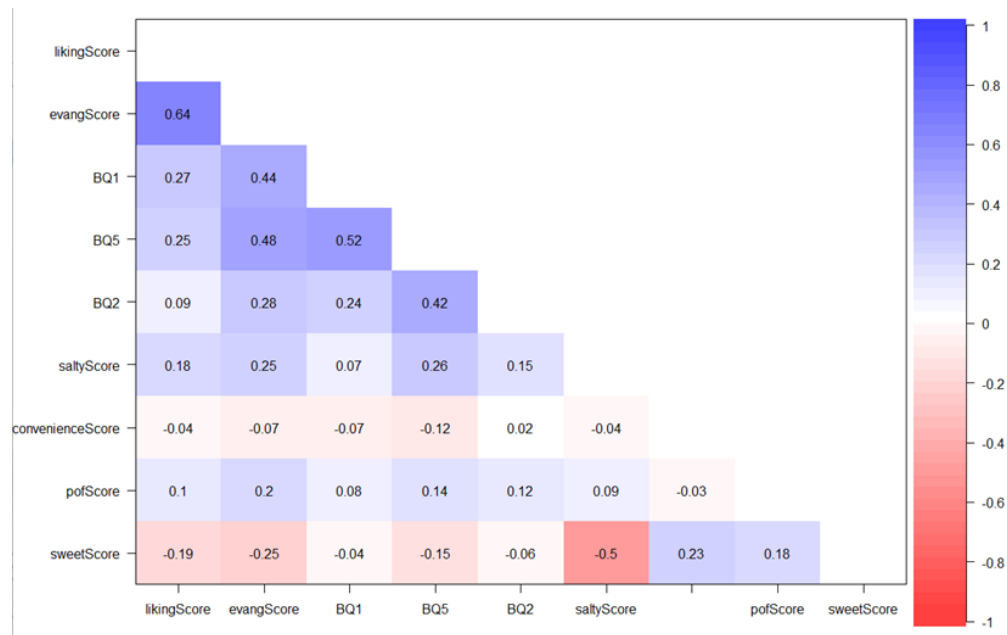
Table 5. The list of the relevant variables, hypotheses, the types of validity evidence

Items	My hypotheses	Types of validity evidence
(BQ1) number of pickled products bought in stores or online	people who like pickled products would tend to purchase pickle products frequently.	Concurrent
(BQ5) how much you would be willing to pay for a pickled product	people who like pickled products would be willing to pay for a pickled product more than those who do not.	Concurrent
(BQ2) how many social media posts about pickled products	people who feel the need to evangelize about the benefits of pickles would tend to make posts	Concurrent

	about pickles on their social media account frequently.	
(saltyScore) liking salty food scale scores	pickle liking would be negatively or weekly associated with salty-food liking.	Discriminant
(convenienceScore) convenience orientation scale scores	some people who like pickles may prefer the foods that can be easy to make/eat, as a pickle is also such a type of food.	Convergent and discriminant
(pofScore) power of food scale scores	people who like foods themselves very much may show a strong preference for pickles as well if pickles are one of their favorites.	Convergent and discriminant
(sweetScore) sweet tooth scale scores	people who strongly like sweet foods may show a strong preference for pickles because some types of pickles are very sweet.	Convergent and discriminant

The correlations between the variables are described in the following figure.

Figure 2. Correlations between relevant variables



The results can be summarized as follows.

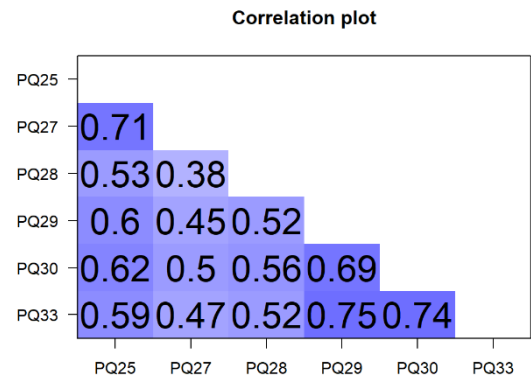
- (1) The correlation between the liking and evangelizing subscales was positive and strong ($r = .64$), which was far larger than correlations with any other variables. It is strong, convergent/discriminant validity evidence for the pickle fanaticism subscales.
- (2) The moderate levels of correlations were observed between likingScore and BQ1 ($r = .27$), likingScore and BQ5 ($r = .25$), and evangScore and BQ2 ($r = .28$), which are concurrent validity evidence for the pickle fanaticism subscales.
- (3) The weak/close-to-zero levels of correlations were observed between likingScore and saltyScore ($r = .18$), convenienceScore and likingScore ($r = -.04$), convenienceScore and evangScore ($r = -.07$), pofScore and likingScore ($r = .1$), indicating the convergent and discriminant validity evidence for the pickle fanaticism subscales.

Overall, this correlation analysis provided some validity evidence based on relations to other variables. However, it is difficult to say that this evidence is sufficient and collective for the two pickle fanaticism subscales. The predictive power of the pickle-liking subscale for its criterion variables was rather weak, indicating the weak concurrent validity for the subscale. Also, the sign and relative size of some correlations were not consistent with theoretical expectations. For instance, the sweet tooth scale has negative and weak-to-moderate correlations with the two pickle fanaticism subscales ($r = -.19$ for likingScore and $r = -.25$ for evangScore). Moreover, the sample turned out not to include any respondents who had a high level of pickle liking. Considering the purpose of this scale (i.e., identifying people with high pickle fanaticism), I should say that the validity evidence obtained from this sample is limited.

VII. Evaluation of Bias and Fairness

Test bias refers to the systematic difference in test scores between two groups even though the two groups have equal levels of the construct being measured by the test. It may occur when construct-irrelevant variance affects responses from one specific group only. For example, when a group of non-native English speakers take the GRE math test, their average test scores can be lower on average than a group of native English speakers because of their gap in English skills. On the other hand, fairness is more relevant to how to use the test scores. If a government agency advertises that they recruit people good at mathematics but evaluates their mathematical skills based on their scores on the GRE math test, a fairness issue may occur. For the pickle fanaticism scale, the fairness issue was less likely to occur because this scale is developed for a private company to find people having a high level of pickle fanaticism and to hire them as social marketers for their products. However, there were two potential issues of test bias with the scale, particularly for the subscale for pickle evangelism.

At first, items 25 ('My family should eat pickles regularly') and 27 ('Everyone should eat pickles') seem to be vulnerable to construct-irrelevant variance. These items talk about how other people should do, beyond respondents' feelings on pickled items. Unlike other items for the subscale of pickle evangelism, these items cannot be endorsed unless someone thinks he or she has the right/expertise to tell others about what to eat. Thus, even though two groups of people feel the strong need to evangelize about the benefits of pickles in the same way, one group may not endorse these items if they think they should not boss around others. As shown in the Figure on the right side, the correlations between the two items ($r = .71$) were higher than those with other items (less than $r = .62$), suggesting the possibility of bias for the two items.



Second, item 25 ('My family should eat pickles regularly') also might have another issue of potential bias. The target to evangelize in this item was too specific ('family') to measure the general feeling to evangelize about pickled items. As a result, this item could be endorsed only if people currently have a positive relationship with their families. If one's family all died or he/she does not like his/her family, he/she would not endorse this item, even though they have a strong feeling to evangelize about the benefits of pickled items. Unfortunately, I couldn't find empirical evidence to support this hypothesis because the old questionnaire did not ask respondents about their relationship with families.

To further investigate the two potential issues above, it is necessary to collect data for criterion variables that can be used to check the predictive relevance of the items. I proposed to add to the survey questionnaire the items of interpersonal tolerance scale and an item for respondents' relationship with their family and to conduct the survey again to obtain data. Once the data is newly collected, I will apply the confirmatory factor analysis to obtain correlations between the items of pickle evangelism and the factor of interpersonal tolerance. If the correlations between items 25 and 27 and the interpersonal tolerance are substantially higher than those between the rest of the pickle evangelism items and the interpersonal tolerance, it may imply that items 25 and 27 may suffer from the construct-irrelevant variance, thereby causing the test bias. Also, I will identify people having a high level of pickle evangelism and classify those people into two groups: people having a positive relationship with their family and people not having such a relationship. If the former group's average score of pickle evangelism is significantly higher than the latter group's average score, it also will serve as empirical evidence to show the possibility of test bias in the pickle evangelism scale.

VIII. Overall Evaluation of Evidence and Recommendations

According to the request from EvilCorp, I conducted a validation study for the pickle fanaticism scale. In Section 2, I reviewed the quality of 33 candidate items for pickle fanaticism from a viewpoint of a domain expert to find validity evidence based on test content for the items. Also, I examined the validity evidence based on response process by carrying out think-aloud interviews and retrospective probes with two participants. By this procedure, I identified 23 problematic items and suggested removing 13 items and revising 10 items. In Section 3, I conducted the descriptive item analysis for the 33 items of pickle fanaticism to obtain three types of validity evidence: evidence based on content validity, evidence based on cognitive response process, and evidence based on the internal structure. Again, I found out that 25 items were problematic and suggested removing nine items among them and revising the rest of the items. In Section 4, I conducted exploratory factor analysis (EFA) for the 11 items of pickle fanaticism to obtain validity evidence based on internal structure. I found out that the two-factor solution was statistically supported by data and theoretically plausible as well. Consequently, a set of items 11, 17, 20, 21, 22, and 25 was determined as a subscale of factor 1 ('pickle liking') and a set of items 27, 28, 29, 30, and 33 as a subscale of factor 2('pickle evangelism'). In Section 5, I conducted reliability analysis for each subscale to find validity evidence based on internal structure. The results showed that the (internal consistency) reliability estimates α of both subscales were greater than .8, which could be considered acceptable unless they are used for clinical purposes. In Section 6, I examined the correlations between the two subscale scores and other relevant variables to obtain validity evidence based on relations to other variables. I found out that the two subscale scores were associated with other variables as theoretically expected in most cases. In Section 7, I reviewed the potential issue of bias and fairness for the two subscales, which are relevant with validity evidence based on internal

structure and on consequence of testing. The two items of the subscale of pickle evangelism might cause test bias, but I couldn't find the potential fairness issue.

However, the 11 items of pickle fanaticism were not free from limitations. At first, the Think-aloud interviews revealed that the Likert scale used in the items might confuse respondents not having any preference for pickled items. In this case, for instance, they could choose either “Strongly disagree” or “Neither agree nor disagree” for item 22 (“I like pickles a lot”). Second, only one domain expert reviewed the quality of items, so that the problems of item content might not have been sufficiently detected. Likewise, as think-aloud interviews were conducted with just two people who were commonly male in their early thirties, it would be difficult to say that the 11 items of pickle fanaticism are sufficiently reviewed in terms of the general population's cognitive process. Particularly, the data did not include any respondents having a high level of pickle liking, which means it is still unknown how people respond to the items when they extremely like pickles. Third, the subscales of pickle fanaticism had rather weak prediction power for their criterion variables and their association patterns were against theories in some cases. Lastly, a couple of items for the pickle evangelism subscale seemed to be affected by external factors that were irrelevant to pickle fanaticism, which could be a source of test bias.

Despite the limitations, overall, the 11 items for the pickle fanaticism scale are validated and finally selected according to the scientific procedure based on psychometrics. The five types of validity evidence obtained from the study support the interpretation and uses of their scores for identifying those who have a high level of pickle fanaticism to some degree. In the qualitative item review and descriptive item analysis, no substantial issues were found from those items. EFA results justified the usage of 11 items as the two subscales of the pickle fanaticism, and reliability analysis showed each subscale had an acceptable level of internal consistency. The correlation analysis showed that the subscale scores were associated with other criterion variables as theoretically expected in most cases.

To improve the scale of pickle fanaticism, we may conduct another validation study for its 11 items. In this study, ‘multiple’ experts should review the item content and conduct think-aloud interviews with a large number of people from ‘diverse’ populations. Above all, it would be important to investigate whether the Likert scale can be interpreted consistently across people having different levels of pickle fanaticism. After revising the items based on their feedbacks, a large sample should be collected such that it can include people having a high level of pickle fanaticism. The new survey questionnaire should additionally include the items of interpersonal tolerance scale and an item for respondents’ relationship with their family to further examine the potential bias of the two items of the pickle evangelism subscale. It would be also worthwhile to refine the items for criterion variables so as to test the predictive relationship between those variables and the pickle fanaticism subscales.

References

- Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. *British Journal of Statistical Psychology*, 4(1), 1–2. [https://doi.org/https://doi.org/10.1111/j.2044-8317.1951.tb00299.x](https://doi.org/10.1111/j.2044-8317.1951.tb00299.x)
- Edwards, A. L. (1957). Techniques of attitude scale construction. In *Techniques of attitude scale construction*. Appleton-Century-Crofts.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151. <https://doi.org/10.1177/001316446002000116>
- Warne, R. T., & Larsen, R. (2014). Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. *Psychological Test and Assessment Modeling*, 56(1), 104–123.