Received: 6 March 2023 | Accepted: 16 April 2023

DOI: 10.1002/agm2.12254

Aging Medicine Open Access WILEY

ORIGINAL ARTICLE

Sensitivity and specificity of conventional and new face validation in determining the incomprehensible items by older people: Empirical evidence of testing 106 quality-of-life items

Simon Ching Lam[1] | Lorna Kwai Ping Suen[1] | Emma Yun-Zhi Huang[2] | Eliza Mi Ling Wong[1] | Daphne Sze Ki Cheung[3] | Rick Yiu Cho Kwan[1]

[1]School of Nursing, Tung Wah College, Hong Kong SAR, China

[2]Nutrilite Health Institute, Shanghai, China

[3]School of Nursing, The Hong Kong Polytechnic University, Hong Kong SAR, China

Correspondence
Simon Ching Lam, School of Nursing, Tung Wah College, Hong Kong SAR, China.
Email: simlc@alumni.cuhk.net; simonlam@twc.edu.hk

Emma Yun-Zhi Huang, Nutrilite Health Institute, Building 6, No.720, Cailun Road, Pudong New Area, Shanghai, China.
Email: huangyunzhiemma@sina.com

Abstract

Objective: This methodological research aimed to investigate and compare the sensitivity and specificity of conventional and new face validation in identifying incomprehensible items empirically.

Methods: A purposive sample of 15 older people living in three residential care homes (RCHs) in Hong Kong was used to evaluate a newly developed 106 items covering seven quality-of-life dimensions. The abbreviated Mental Test (Hong Kong version; AMT) was used as a screening tool for excluding those with impaired cognition. The interview was audiotaped, and incomprehensible items were identified by the research panel accordingly (served as the gold standard). The socio-demographics of the respondents were described. Understandability (yes/no, conventional face validation method) and interpretability (4-point Likert scale, new method) were compared and used to compute the *Kappa* value (representing chance agreement), sensitivity, and specificity analysis.

Results: Fifteen older people were interviewed and responded to the structured interview of 106 items regarding understandability and interpretability. 61 items (57%) obtained 100% positive understandability while only 35 items (33%) obtained 100% correct interpretability.

The *Kappa* coefficient was 0.388 ($P < 0.001$) of the chance agreement between understandability and interpretability. The panel confirmed that 32% of items required revision (i.e., incomprehensible items). The false negative rate of using the conventional approach was up to 70.59% while both the false positive and negative rates of using the new approach were low (0%–5.88%).

Conclusion: This empirical evidence indicated that the conventional approach of face validation for checking incomprehensible items by older people encountered a high false negative rate. On the contrary, the new approach was recommended because it demonstrated high sensitivity and specificity and low false positive and negative rates in identifying incomprehensible items.

KEYWORDS
empirical evidence, face validation, interpretability, older people, understandability

1 | BACKGROUND

Literature has raised up an issue about the value of face validation in the psychometric testing of instruments.[1–3] Face validation is a controversial issue because researchers define it as pertaining to the superficial examination of an instrument, by checking the understandability of respondents and acceptability of both respondents and administrators by a yes or no response.[1,4–6] Collectively, the meaningful argument has been more focused on the function than its power in psychometric validation. Several useful and unique functions, namely enhancing the motivation and cooperation of respondents, reducing dissatisfaction among stakeholders, and increasing the acceptability of the findings by policymakers, were frequently mentioned.[7–11] However, the conventional testing approach has its shortcomings when applied to older people or people with low educational levels (e.g., child).[12] Particularly, it is questionable whether the conventional face validation method can identify items that are difficult to comprehend by the target population.

1.1 | Face validity and conventional approach of conducting face validity

Testing face validity refers to "whether the instrument looks like it is measuring the target construct."[13] Given that such property pertains to how the stakeholders of the instrument perceive it, face validity should be judged by them and not by experts in the field.[14,15]

The mainstream school of thought recommends assessing face validity by nominal scale (i.e., all or none, yes or no) due to the lack of standards for judging it or for determining how much of it an instrument has.[14] Concerning the procedure of conducting face validity tests, the literature recommends testing the understandability of a target population who would like to respond to the questions presented in an instrument (e.g. do you understand the question/item?), and the acceptability of both the respondents and administrators who will potentially administer the instrument.[3,14,16–18]

1.2 | Theoretical background of the new approach

To our knowledge, there was limited research that explores whether the conventional approach (checking understandability) can achieve the purpose of testing (identification of incomprehensible items). However, correct interpretation of the items by the designated target population is the prerequisite of a valid instrument, particularly for a self-reported measurement.[14,16] The new approach of face validation was informed by the theoretical framework of "symbolic interactionist theory" (refer to the supplementary material) from Herbert Blumer.[18] According to this framework on question-answer behavior, an older respondent goes through the process of "decoding the question" and "encoding the answer." Generally, the capability of an older respondent to understand would affect how he or she decodes a question; the capability of an older respondent to accept would affect how he or she encodes the answer.[18] Therefore, the "correct interpretation" of the items from the respondents would be the ultimate goal of the face validation instead of their understandability because the former verified the latter.[4,5,7-9,14–16] The "correct interpretation" was able to be assessed by two approaches. First, with reference to a technique introduced by Nuckols on testing questions,[15] the respondents would be asked to rephrase the questions in their own words and keep the meaning as close to the original question as possible. Second, some elders may find it difficult to rephrase the questions because of limited education or a shortage of wording. They would be encouraged to describe their daily experience or life scenario to reflect their interpretation of the meaning of questions.[15,16] The researchers could use their responses and code them into one of four categories: fully correct, generally correct (less than one part of the meaning is altered or omitted), partially wrong (the respondent understood the intent), and completely wrong on interpretation or cannot be rephrased. Seeking their suggestions on wordings, terminology, or the structure of questions is necessary once the response is identified as "partially wrong" or "completely wrong on interpretation or cannot be rephrased."[15,16] Furthermore, it is suggested that conversations for conducting face validity should be digitally recorded and transcribed verbatim, if necessary, which can be conveniently used in post-interview clarifications through a panel discussion. In a word, the correct interpretation is the ultimate goal of a good item in the validation of an instrument.[11,14–16]

The previous literature has discussed and identified the shortcomings of the conventional approach of face validation and recommended a new approach to enhance the validity of testing results.[1,3,12] However, those discussions (i.e., comparison of conventional and new approaches) were purely theoretical. As we know from the existing literature, there is no empirical research demonstrating the evidence of such an argument. Therefore, the current study aimed to compare conventional and new approaches to face validation empirically with reference to the gold standard.

2 | METHODS

2.1 | Study design

A cross-sectional methodological study was conducted, and an interview lasted for approximately 90 min for each participant was audiotaped. Each participant was asked for understandability and

interpretability for each item of the designated instrument (i.e., newly developed 106 items for assessing the quality of life).

2.2 | Instruments

2.2.1 | Quality of life scale for residential care home elders (QOL-RCHE)

The QOL-RCHE is the first situation-specific quality-of-life instrument developed for older Chinese people in RCHs. It contains 106 items covering seven dimensions (i.e., good living, physical well-being, spiritual well-being, psychological well-being, social well-being, satisfaction with the RCH environment, and self-worth) in the phase 1 study: Item generation.[19] These items were newly constructed through the information from the literature, then developed by a focus group interview with experts, and individual interviews with 30 residents.[19] The items have previously gone through content validation.

2.2.2 | Abbreviated mental test (Hong Kong version; AMT)

The AMT was used for checking the cognitive status of our older participants. It was widely used to screen impaired cognitive function of older people in Hong Kong and the best cutoff value was 7.[19–21] When applying AMT in older people in community and nursing homes, the psychometric properties were established satisfactorily and a sensitivity of 92.3% and specificity of 87.1% was attained.[20]

2.3 | Samples and sample size

A purposive sampling consisting of 15 elders living in three residential care homes (RCHs) in Hong Kong was conducted. The sample size estimation for face validation does not involve power and effect size. The systematic review indicated no consent for sample size (range: 3–603, median = 20).[22] However, about 15 to 20 samples are generally accepted for instrument development[13,15] and such a suggestion has been tested in several recent studies.[23–26]

All participants involved in our study met the following inclusion criteria: (1) they were able to articulate their experiences in Cantonese; (2) cognitively competent (AMT ≥ 7);[21] (3) did not have any psychiatric illness; (4) lived in the RCH for more than 6 months.[27] The approach of maximum variation in sample selection was purposefully adopted.[19] The researcher intended to select older people with purpose who were from different educational (illiterate to tertiary educated) and cognitive levels (AMT = 7–10). It was generally believed that these characteristics determined the interpretation and understandability of items.[19,23,27]

2.4 | Study procedure and gold standard

The participants signed a consent form and completed a simple demographic data sheet (i.e., age, educational level, and types of RCHs). A structured interview was used to conduct the face validation test. The research assistant conducted the conventional approach prior to the new approach. Each interview lasted for approximately 90 min. The interview process was recorded using a digital voice recorder for subsequent verification of any uncertain responses. The research panel consisted of experts in instrument development, and academic staff with doctorates in health care and language, which determined the incomprehensible items based on the recorded conversation. That decision served as the gold standard to keep, revise, or discard the items.

2.5 | Statistical analysis

Data were collected and analyzed using Excel and SPSS (version 24). Percentages were used to report categorical variables.

For conventional and new approaches of face validation, to determine incomprehensible items, it was recommended if there was over 20% of participants did not understand the particular item or incorrectly interpreted it.[15,19] Therefore, the number of items that were identified as incomprehensible was computed by the conventional approach and new approach respectively. The chance agreement between these two approaches was computed through the *Kappa statistic*. A value of above 80% was considered as excellent agreement; above 60% substantial levels of agreement; 40% to 60% moderate; and below 40% poor to fair agreement.[28]

Then, the results of understandability and interpretability were compared with the gold standard, in other words, which were the conventional and new approaches compared with the decision of the research panel, respectively, regarding incomprehensible items. The 2 × 2 contingency table on the overall performance of conventional and new approaches was constructed against the gold standard. The sensitivity, specificity, accuracy, false positive rate, and false negative rate were then illustrated. p value was set at < 0.050 as significance.

Lastly, the narrative interpretations from participants were illustrated for a better understanding of how the wrong interpretation occurred.

3 | RESULTS

3.1 | Characteristics of older people participants

Fifteen elders from three different RCHs were interviewed for testing face validation. The participants had diverse age ranges (mean = 80.3; ranged from 73 to 86), educational backgrounds (illiterate to tertiary educated), religious beliefs, length of stay (range from 3 months to 10 years), and level of physical independence

(ranged from bed-bound to physical independence). By using AMT as a screening tool, all participants were of normal cognition. However, the AMT score could be a reference to indicate the level of cognitive ability, which a higher AMT score represents better cognitive ability. The researcher recruited the elders with different AMT scores (i.e., 7, 8, 9, and 10). Table 1 presents the socio-demographic characteristics of the participants.

3.2 | Understandability of items using the conventional approach

Among 106 items, the participants reported 100% understandability (i.e., all participants can understand it) on 61 (57%) items. The range of understandability across the items was 67% to 100% (data not shown). Table 2 illustrated the partial items for the face validation test during the structural interview for checking the understandability, interpretability, and decision of the research panel of 51 selected items. By the use of the conventional method, only 11 items (10.37% of 106 items) obtained poor understandability (positive for identifying as incomprehensible items ranged from 20.0% to 33.3%).

3.3 | Interpretability of items using the new approach

As for adopting the new approach, nevertheless, only 35 items (33%) obtained a 100% correct interpretation (i.e., all participants can interpret it correctly). Most of the items (67%) could not be fully or generally rephrased or meaning correctly explained by the participants. By the use of the new method, about 32 items (30.19% of 106 items) obtained poor interpretability (positive for identifying as incomprehensible items ranged from 20.0% to 60.0%). For some particular items (i.e., So1.12 & V2.2), over half of the participants failed to interpret these items correctly (Table 2).

3.4 | Chance agreement, sensitivity, and specificity of the two approaches

The chance agreement between conventional and new approaches was low (*Kappa coefficient* = 0.388, $P < 0.001$).

When the number of incomprehensible items ($n = 34$, 32.07% of total items) identified by the research panel was taken into account, the conventional and new approaches were compared accordingly (Tables 3–5). The *Kappa coefficient* was 0.361 ($P < 0.001$) between the conventional and the gold standard. The sensitivity and specificity of the conventional approach for identifying the incomprehensible items were 29.41% and 98.61%, respectively. The false negative rate was up to 70.59%. The overall accuracy was only 76.42%.

The *Kappa coefficient* was 0.956 ($P < 0.001$) between the new and the gold standard, indicating a high chance of agreement. The sensitivity and specificity of the new approach for identifying the incomprehensible items were 94.12% and 100%, respectively. The false positive and negative rate was very low. The overall accuracy was up to 98.11%.

TABLE 1 Socio-demographic characteristics of the participants ($n = 15$).

Code	Age	Gender	Religion	Types of RCHs	Educational level	Total Length of stay RCH (year)
01	79	M	Nil	Subvented	Primary	0.5
02	67	M	Christian	Subvented	Tertiary	7.5
03	85	F	Catholic	Subvented	Primary	2.5
04	86	F	Catholic	Subvented	Primary	10
05	84	F	Catholic	Private	Primary	3.5
06	79	F	Nil	Private	Secondary	0.6
07	90	M	Nil	Private	Illiterate	3
08	75	F	Buddhism	Private	Tertiary	1
09	81	M	Nil	Private	Secondary	0.6
10	81	M	Buddhism	Private	Primary	2.6
11	83	F	Buddhism	Private	Illiterate	0.5
12	73	M	Christian	Private	Secondary	7.5
13	81	M	Buddhism	Subvented	Primary	3
14	78	F	Ancestor Worship	Private	Illiterate	3
15	83	M	Nil	Private	Illiterate	1.5

Note: Subvented, included subvented RCH or nursing homes operated by non-government organization.

Private, private nursing home.

TABLE 2 Understandability, interpretability, and decision of the research panel among quality-of-life items ($n = 15$ older people).

51 out of 106 items	Und./Interp. (%)	Decision from the Panel	Action	No	Incom
P1.3 I do not think of my health condition as others' burden.	0.00/ 26.67	Incomprehensible	Revised	1	1
P1.4 The RCH allows me to take care of my daily life where I can manage.	0.00/ 13.33	Ok	Revised	2	
P1.6 With my present health condition, I can move around with ease in the RCH.	6.67/ 6.67	Ok	Revised	3	
P1.7 I think I can live independently with the help of the RCH.	13.33/ 33.33	Incomprehensible	Revised	4	2
S1.1 Spiritual beliefs (like religion, ancestral worship, God and fate) provide me the energy to face RCH life.	15.38/23.08[a]	Incomprehensible	Revised	5	3
S1.2 Spiritual beliefs (like religion, ancestral worship, God and fate) make my mind calm in the RCH.	7.69/23.08[a]	Incomprehensible	Revised	6	4
S1.3 Spiritual beliefs (like religion, ancestral worship, God and fate) make me feel peaceful in the RCH.	7.69/23.08[a]	Incomprehensible	Revised	7	5
S1.4 At the RCH, I can pursue my spiritual activities (like ancestral worship and religious worship).	7.69/25.00[a]	Incomprehensible	Revised	8	6
S2.1 I think I accomplish most of the important things in life with satisfaction.	13.33/20.00	Incomprehensible	Revised	9	7
S2.2 I think I still have regrets in my life.	13.33/6.67	Ok	Revised	10	
S2.3 At the RCH, I find my life meaningful.	13.33/13.33	Ok	Revised	11	
S3.1 This RCH can help to arrange my end-of-life matters in advance.	6.67/6.67	Ok	Revised	12	
S3.2 At the RCH, my last will is respected.	6.67/13.33	Ok	Revised	13	
Ps1.4 I feel depressed living in RCH.	13.33/40.00	Incomprehensible	Revised	14	8
Ps1.5 I feel anxious living in RCH.	20.00/40.00	Incomprehensible	Revised	15	9
Ps1.7 I feel sorrowful living in RCH.	6.67/40.00	Incomprehensible	Revised	16	10
Ps2.1 I think I have adjusted to my RCH life.	0.00/6.67	Ok	Revised	17	
Ps2.2 I think relocating to RCH is a correct choice.	0.00/0.00	Ok	Revised	18	
Ps2.3 I think relocating to RCH makes me feel abandoned.	0.00/0.00	Ok	Revised	19	
Ps2.4 I think relocating to RCH is something I would not reveal to others.	6.67/20.00	Incomprehensible	Revised	20	11
Ps2.5 I think relocating to RCH makes me lose face.	6.67/6.67	Ok	Revised	21	
Ps2.6 I think relocating to RCH is an embarrassing thing to tell others.	20.00/26.67	Incomprehensible	Deleted	22	12
Ps2.7 I do not need to worry about causing burden to my family after relocating to the RCH.	0.00/13.33	Incomprehensible	Revised	23	13
Ps2.8 Relocating to RCH has solved the problems I used to face at home.	0.00/13.33	Ok	Revised	24	
Ps3.1 I face what I have to face the RCH life with a contented mind.	0.00/7.14[a]	Ok	Revised	25	
Ps3.2 I feel at ease to accept what I have to face in RCH life.	13.33/20.00	Incomprehensible	Revised	26	14
Ps3.3 I reconcile myself to what I have to face in RCH life.	20.00/20.00	Incomprehensible	Revised	27	15
So1.4 At the RCH, I am very distant with other residents.	20.00/20.00	Incomprehensible	Revised	28	16
So1.10 I feel uncomfortable about getting into conflicts with the staff.	0.00/35.71[a]	Incomprehensible	Revised	29	17
So1.11 The RCH environment helps me build good relationships with the staff and other residents.	0.00/35.71[a]	Incomprehensible	Deleted	30	18
So1.12 At the RCH, no one (like staff and residents) has concern for me here.	33.33/50.00[a]	Incomprehensible	Revised	31	19
So3.2 The RCH offers me chance to participate in community activities.	0.00/33.33	Incomprehensible	Revised	32	20
So3.3 At the RCH, I think I am disconnected from society.	13.33/13.33	Ok	Revised	33	
So4.1 At the RCH, the activities that I participate in give me good pastimes.	0.00/6.67	Ok	Revised	34	
So4.2 Activities in the RCH add color to my daily life.	13.33/20.00	Incomprehensible	Revised	35	21
So4.3 At the RCH, I can help others that gives my life a sense of meaning.	6.67/20.00	Incomprehensible	Revised	36	22
So4.4 I am satisfied with my assigned tasks from the RCH.	13.33/26.67	Incomprehensible	Revised	37	23

TABLE 2 (Continued)

51 out of 106 items	Und./Interp. (%)	Decision from the Panel	Action	No	Incom
Sa1.2 For my current physical condition, the medical support provided by the RCH gives me a sense of safety.	20.00/20.00	Incomprehensible	Revised	38	24
Sa1.3 At the RCH, my money and belongings are secured.	6.67/13.33	Ok	Revised	39	
Sa2.10 Residents with certain physical condition and caring requirements make me feel disgusted.	20.00/40.00	Incomprehensible	Revised	40	25
V1.1 At the RCH, I can still keep my favorite habits.	6.67/13.33	Ok	Revised	41	
V1.3 At the RCH, my feelings and needs are valued by the staff.	13.33/26.67	Incomprehensible	Revised	42	26
V2.1 At the RCH, I think that I am being respected.	0.00/13.33	Ok	Revised	43	
V2.2 At the RCH, I can uphold my dignity.	33.33/60.00	Incomprehensible	Revised	44	27
V3.1 The RCH offers me adequate autonomy to decide the details of my daily life.	28.57/28.57[a]	Incomprehensible	Revised	45	28
V3.2 I can still have sufficient choices on various issues of RCH life.	7.14/21.43[a]	Incomprehensible	Revised	46	29
V3.3 RCH regulations do not significantly restrict my living.	0.00/21.43[a]	Incomprehensible	Revised	47	30
V4.1 At the RCH, I can still keep my privacy.	14.26/14.26[a]	Ok	Revised	48	
V4.2 When necessary, I have a place with no one else around to deal with my emotions.	21.43/46.15[a]	Incomprehensible	Revised	49	31
V4.3 The RCH can provide me a private room to catch up with my family.	0.00/28.57[a]	Incomprehensible	Revised	50	32
V4.8 At the RCH, my personal data are kept confidential.	7.14/35.71[a]	Incomprehensible	Revised	51	33

Note: Und./Interp. (%), the percentage of participants who did not understand the item (i.e., positive for the incomprehensible item)/the percentage of participants who incorrectly interpret the item (i.e., positive for the incomprehensible item).

Underline, indication of revision.

Incom, list of incomprehensible items in order; No, list of 51 items in order.

[a]Some participants discontinued the interview or refused to respond to some items.

During the interview, it was observed that the respondents might wrongly interpret the item, even expressing good understanding of it. For instance, the wording of "the issues should be kept secret" of item Ps2.4 was interpreted as "making a decision by myself" or "making a complaint of or suing the other residents." Another example was the wrong emphasis on the focus of the item. The item So1.10 "I feel uncomfortable about getting into conflicts with the staff" focused on whether the residents were concerned about the conflicts with staff. All respondents expressed that they understood this item but over 35% of them put the focus on whether such conflict occurred. Table 6 presents some narrative comments from the respondents during the interview. The respondents commented that lengthy statements, statements with passive voice, and medical or literary terms might increase the difficulty in comprehension.

4 | DISCUSSION

The present study provided the first empirical evidence on the critique of the conventional approach of face validation for identifying incomprehensible items in the process of psychometric testing. The current results echoed the literature that the conventional method (i.e., checking understandability) was unable to identify the problematic items because of low sensitivity and chance agreement and high false negative rate.[3,12] More importantly, the current study added value to the face validation in identifying incomprehensible items through the introduction of the new method, i.e., checking the interpretability. Such initiative vitalized the function of face validation in the field of psychometric testing because incomprehensible items compromised both validity and reliability.

4.1 | Shortcomings of the conventional method

The current results indicated that respondents tended to endorse their good understanding of the item statements.[1,5,15,16,26] Although older people were fully informed on the sequence of the two parts of the interview (i.e., conventional first and then new approach), they would still express their good understanding of the given items and claimed that they did not understand the terms or statement afterwards when checking their interpretability (e.g., P1.3 "health burden," P1.4 "management of daily life," Ps2.8 "family need of relocation in RCH," V2.1 "being respected," V4.3 "provision of a private room for meeting"). This may be one of the reasons contributing to the great discrepancy between understandability and interpretability in identifying incomprehensible items under the process of face validation.

Second, most of the respondents would unintentionally misinterpret the terms of the item or the entire items (e.g., S1.3 "feeling peaceful," S2.1 "accomplishment of important things," Ps1.4 "feeling depressed," Ps1.5 "feeling anxious," Ps2.5 "losing face for

TABLE 3 Cross-tabulation of the results of the conventional approach against the gold standard for indicating the incomprehensible items (item number = 106).

Conventional method to identify the incomprehensible item (Understandability)	The research panel's decision for incomprehensible items (gold standard)		
	Positive, incomprehensible items	Negative, comprehensible items	Totals
Positive (not understandable items)	10	1	11
Negative (understandable items)	24	71	95
Totals	34	72	106

TABLE 4 Cross-tabulation of the results of the new approach against the gold standard for indicating the incomprehensible items (item number = 106).

New method to identify the incomprehensible item (Interpretability)	The research panel's decision for incomprehensible items (gold standard)		
	Positive, incomprehensible items	Negative, comprehensible items	Totals
Positive (not interpretable items)	32	0	32
Negative (interpretable items)	2	72	74
Totals	34	72	106

TABLE 5 Results of the conventional approach for identifying incomprehensible items (item number = 106).

Items	Percentage (%) or Kappa
Items indicated for revision through the conventional approach	10.37%
Items indicated for revision through the research panel	32.07%
Kappa coefficient	0.341, $P < 0.001$
Sensitivity	29.41%
Specificity	98.61%
Accuracy	76.42%
False positive rate (FPR)	1.39%
False negative rate (FNR)	70.59%

Note: Accuracy = (TP + TN)/(P + N).

TABLE 6 Results of the new approach for identifying incomprehensible items (item number = 106).

Items	Percentage (%)
Items indicated for revision through the new approach	30.19%
Items indicated for revision through the research panel	32.07%
Kappa coefficient	0.956, $P < 0.001$
Sensitivity	94.12%
Specificity	100%
Accuracy	98.11%
False positive rate (FPR)	0.00%
False negative rate (FNR)	5.88%

relocation," Ps2.6 "embarrassing to be relocated," Ps2.7 "reduce the burden to my family after relocation," So1.10 "feeling uncomfortable in the conflicts with staff"). Last, some respondents addressed the wrong focus of the item (e.g., P1.7 "live independently with the help of RCH," So1.10 "feeling uncomfortable in the conflicts with staff"). Given the limitation of the self-reported scale, the researcher has no chance to clarify the interpretation of the items by the respondents. Misinterpreting the items would certainly worsen the reliability and validity of measurement or assessment. Therefore, checking the interpretability of the items of a self-reported instrument, namely

rephrasing the items or description of a similar daily scenario, would eventually facilitate the identification of incomprehensible questions, which was similar to the concept of "cognitive debriefing".[29] Both the high sensitivity and specificity demonstrated sufficient evidence to this claim in the current study.

4.2 | Value of the conventional method

The high false negative rate (low sensitivity) indicated that the capability of the conventional approach to accurately assess the presence of the condition (i.e., incomprehensible item) for the items was poor (known as a high missing rate). Contrarily, the conventional approach obtained a low false positive rate, which meant the ability of this approach to accurately assess the absence of the condition was good. In other words, older people respondents who did not understand the item would be unable to interpret it, which indicated an incomprehensible item.

4.3 | Limitations

Some limitations deserved discussion. This study demonstrated the shortcomings of the conventional method for face validation. However, the respondents consisted of mainly low educated or even illiterate, which may decrease the chance of correct interpretation of the testing items. Therefore, the disadvantage of the conventional approach may be magnified. Similarly, this study adopted the 106 items that were newly developed, preliminary, and not yet tested. It is anticipated that more items were poorly devised. Thus, the result might not be generalized to those validated scales.

5 | CONCLUSION

This empirical study demonstrated the shortcoming of the conventional approach to face validation and recommended the new approach for checking the interpretability of the older people to the items. This new approach was constructed based on the symbolic interactionist theory and Nuckols' technique on testing questions. Through the empirical testing on older people, the chance agreement between the conventional and new approaches was poor. A high false negative rate of the conventional approach to identifying incomprehensible items was found. Therefore, it was suggested to assess the respondents' interpretation as the approach for face validation.

AUTHOR CONTRIBUTIONS

S.C.L. designed this empirical study and wrote the original manuscript. E.Y.Z.H revised the manuscript and coordinated the submission. L.K.P.S., E.M.L.W., D.S.K.C., and R.Y.C.K. provided a critical review to finalize the draft. All authors provided intellectual content, reviewed the manuscript, and agreed on the submitted version.

ACKNOWLEDGMENTS

The authors thank all the older people respondents for their participation in our study. The first author also wants to thank Prof. Diana Lee and Prof. Doris Yu for supervising his 3-phase PhD project, of which the current data come from a part of the phase 2 study.

FUNDING INFORMATION

Not Applicable.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets presented in this study are subjected to the following restrictions: Datasets will be available upon reasonable request. Requests to assess the datasets should be directed to Simon Ching Lam, simonlam@twc.edu.hk or simlc@alumni.cuhk.net.

ETHICAL CONSIDERATION

This study received ethical approval from the Survey and Behavioral Research Ethics.

Committee of the Chinese University of Hong Kong (Ref no: SBREC-02092009). All older people participants signed a consent form before the face validation test.

ORCID

Simon Ching Lam https://orcid.org/0000-0002-2982-9192
Lorna Kwai Ping Suen https://orcid.org/0000-0002-0126-6674
Emma Yun-Zhi Huang https://orcid.org/0000-0001-5967-2731
Eliza Mi Ling Wong https://orcid.org/0000-0003-0698-9000
Daphne Sze Ki Cheung https://orcid.org/0000-0001-5651-9352
Rick Yiu Cho Kwan https://orcid.org/0000-0002-4332-780X

REFERENCES

1. Holloway L, Humphrey L, Heron L, et al. Patient-reported outcome measures for systemic lupus erythematosus clinical trials: a review of content validity, face validity and psychometric performance. *Health Qual Life Outcomes*. 2014;12:116. doi:10.1186/s12955-014-0116-1
2. Campbell SM, Hann M, Roland MO, Quayle JA, Shekelle PG. The effect of panel membership and feedback on ratings in a two-round Delphi survey. *Med Care*. 1999;37(9):964-968. doi:10.1097/00005650-199909000-00012
3. Sartori R. Face validity in personality tests: psychometric instruments and projective techniques in comparison. *Qual Quant*. 2009;44(4):749-759. doi:10.1007/s11135-009-9224-0
4. Anastasi A, Urbina S. *Psychological Testing*. 7th ed. Prentice Hall; 1997.
5. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev*. 2004;111(4):1061-1071. doi:10.1037/0033-295x.111.4.1061
6. Allen MJ, Yen WM. *Introduction to Measurement Theory*. 1st ed. Brooks/Cole; 1979.

7. Hogan TP. *Psychological Testing: A Practical Introduction*. 4th ed. John Wiley & Sons; 2019.
8. Kline P. *The Handbook of Psychological Testing*. Routledge; 1993.
9. Royal K. "Face validity" is not a legitimate type of validity evidence! *Am. J. Surg*. 2016;212(5):1026-1027.
10. Mosier CI. A critical examination of the concepts of Face validity. *Educ Psychol Meas*. 1947;7(2):191-205. doi:10.1177/001316444700700201
11. Nevo B. Face validity revisited. *J Educ Meas*. 1985;22(4):287-293. doi:10.1111/j.1745-3984.1985.tb01065.x
12. Newfields T. Challenging the notion of face validity. *Shiken: JALT Testing & Evaluation SIG Newsletter*. 2002;6(3):19 Accessed March 28, 2023. https://hosted.jalt.org/test/new_2.htm
13. Polit DF, Beck CT. *Nursing Research: Generating and Assessing Evidence for Nursing Practice*. 10th ed. Lippincott Williams & Wilkins; 2016.
14. Portney LG. *Foundations of Clinical Research: Applications to Evidence-based Practice*. 3rd ed. FA Davis; 2020.
15. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to their Development and Use*. 5th ed. Oxford University Press; 2015.
16. Lam SC. Universal to standard precautions in disease prevention: preliminary development of compliance scale for clinical nursing. *Int J Nurs Stud*. 2011;48(12):1533-1539. doi:10.1016/j.ijnurstu.2011.06.009
17. Palmieri A, Sorarù G, Lombardi L, et al. Quality of life and motor impairment in ALS: Italian validation of ALSAQ. *Neurol Res*. 2010;32(1):32-40. doi:10.1179/174313209x385734
18. Foddy WH. *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*. Cambridge University Press; 1993.
19. Lam SC. *Development and Validation of a Quality of Life Instrument for Older Chinese People in Residential Care Homes*. Ph.D. Thesis: The Chinese University of Hong Kong, The Nethersole School of Nursing 2015.
20. Lam SC, Wong Y, Woo J. Reliability and validity of the abbreviated mental test (Hong Kong version) in residential care homes. *J Am Geriatr Soc*. 2010;58(11):2255-2257. doi:10.1111/j.1532-5415.2010.03129.x
21. Huang EYZ, Cheung J, Liu JYW, Kwan RYC, Lam SC. Groningen frailty indicator–Chinese (GFI-C) for pre-frailty and frailty assessment among older people living in communities: psychometric properties and diagnostic accuracy. *BMC Geriatr*. 2022;22(788):1-13. doi:10.1186/s12877-022-03437-1
22. Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health Qual Life Outcomes*. 2014;12(2):176. doi:10.1186/s12955-014-0176-2
23. Lam SC, Yeung CCY, Chan JHM, et al. Adaptation of the score for allergic rhinitis in the Chinese population: psychometric properties and diagnostic accuracy. *Int Arch Allergy Immunol*. 2017;173(4):213-224. doi:10.1159/000477727
24. Liu TW, Lam SC, Chung MH, Ho KHM. Adaptation and psychometric testing of the hoarding rating scale (HRS): a self-administered screening scale for epidemiological study in Chinese population. *BMC Psychiatry*. 2020;20(1):1-10. doi:10.1186/s12888-020-02539-7
25. Li S, Kwok SWH, Siu SCN, et al. Development of generic student engagement scale in higher education: an application on healthcare students. *Nurs Open*. 2023;10(3):1545-1555. doi:10.1002/nop2.1405
26. Kwan RYC, Lam SC, Wang SL, Wong AKC, Shi L, Wong FKY. Perception of E-health Technology Scale in Chinese Brief (PETS-C Brief): Translation, item reduction, and psychometric testing. *Digit Health*.2022;8:205520762211260.doi:10.1177/20552076221126055
27. Lee DTF. Perceptions of Hong Kong Chinese elders on adjustment to residential care. *J Interprof Care*. 2001;15(3):235-244. doi:10.1080/13561820120063129
28. Richard LJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174. doi:10.2307/2529310
29. Ploughman M, Austin M, Stefanelli M, Godwin M. Applying cognitive debriefing to pre-test patient-reported outcomes in older people with multiple sclerosis. *Qual Life Res*. 2010;19(4):483-487. doi:10.1007/s11136-010-9602-z

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lam SC, Suen LKP, Huang E-Z, Wong EML, Cheung DSK, Kwan RYC. Sensitivity and specificity of conventional and new face validation in determining the incomprehensible items by older people: Empirical evidence of testing 106 quality-of-life items. *Aging Med*. 2023;6:230-238. doi:10.1002/agm2.12254