

Development and Initial Validation of the Observer-Rated Housing Quality Scale (OHQS) in a Multisite Trial of Housing First

Carol E. Adair, Brianna Kopp, Jennifer Lavoie, Jino Distasio,
Stephen W. Hwang, Aimee Watson, Scott Veldhuizen,
Katherine Chislett, Jijian Voronka, Muznah Ahmad,
Naveed Ahmed, and Paula Goering

ABSTRACT *Quality of housing has been shown to be related to health outcomes, including mental health and well-being, yet “objective” or observer-rated housing quality is rarely measured in housing intervention research. This may be due to a lack of standardized, reliable, and valid housing quality instruments. The objective of this research was to develop and validate the Observer-Rated Housing Quality Scale (OHQS) for use in a multisite trial of a “housing first” intervention for homeless individuals with mental illness. A list of 79 housing unit, building, and neighborhood characteristics was generated from a review of the relevant literature and three focus groups with consumers and housing service providers. The characteristics were then ranked by 47 researchers, consumers, and service providers on perceived importance, generalizability, universality of value, and evidence base. Items were then drafted, scaled (five points, half values allowed), and pretested in seven housing units and with seven raters using cognitive interviewing techniques. The draft scale was piloted in 55 housing units in Toronto and Winnipeg, Canada. Items were rated independently in each unit by two trained research assistants and a housing expert. Data were analyzed using classical psychometric approaches and intraclass correlation coefficients (ICC) for inter-rater reliability. The draft scale consisted of 34 items assessing three domains: the unit, the building, and the neighborhood. Five of 18 unit items and 3 of 7 building items displayed ceiling or floor effects and were adjusted accordingly. Internal consistency was very good (Cronbach’s $\alpha=0.90$ for the unit items, 0.80 for the building items, and 0.92 total (unit and building)). Percent agreement ranged from 89 to 100 % within one response scale value and 67 to 91 % within one half scale value. Inter-rater reliability was also good (ICCs were 0.87 for the unit, 0.85 for the building, and 0.93 for the total scale). Three neighborhood items (e.g., distance to transit) were found to be most efficiently rated using publicly available information. The physical quality of housing can be reliably rated by trained but nonexpert raters using the OHQS. The tool has potential for improved measurement in housing-related health research, including addressing the limitations of self-report, and may also enable documenting the quality of housing that is provided by publicly funded housing programs.*

Adair is with the Departments of Psychiatry and Community Health Sciences, University of Calgary, 1215 – 39 Ave., SW, Calgary, Alberta T2T 2K6, Canada; Kopp is with the Mental Health Commission of Canada, Calgary, Alberta, Canada; Lavoie is with the Department of Criminology, Wilfrid Laurier University, Brantford, Ontario, Canada; Distasio is with the Institute of Urban Studies, University of Winnipeg, Winnipeg, Manitoba, Canada; Hwang, Ahmad, and Ahmed are with the Centre for Research on Inner City Health, St. Michael’s Hospital, Toronto, Ontario, Canada; Watson is with the Mental Health Commission of Canada, Toronto, Ontario, Canada; Veldhuizen and Goering are with the Centre for Addiction and Mental Health, Toronto, Ontario, Canada; Chislett is with the Community Services Department, Regional Municipality of Niagara, St. Catharines, Ontario, Canada; Voronka is with the Department of Humanities, Social Science and Social Justice, University of Toronto, Toronto, Ontario, Canada.

Correspondence: Carol E. Adair, Departments of Psychiatry and Community Health Sciences, University of Calgary, 1215 – 39 Ave., SW, Calgary, Alberta T2T 2K6, Canada. (E-mail: ceadair@ucalgary.ca)

KEYWORDS *Housing quality, Standardized measures, Housing-related health research, Public housing, Homelessness, Mental health*

BACKGROUND and RATIONALE

The physical quality of housing has been associated with many health conditions. Relationships between single exposures and single health outcomes are long established (e.g., residential lead paint and neurodevelopmental toxicity)¹ and continue to be identified (e.g., home heating levels and chronic obstructive pulmonary disease status).² Housing quality, measured by combining multiple characteristics, has also been associated with a range of adverse physical health outcomes, including infectious diseases, respiratory diseases, injuries, cardiovascular diseases,¹ and overall mortality.³ In the past decade the evidence base for housing and health outcomes has been further strengthened through intervention research.⁴ With respect to housing quality and mental health associations, the evidence has emerged more recently but is also accumulating.^{5,6} Examples of single-exposure/outcome associations include the relationships between insufficient daylight and depressive symptoms and between noise levels and psychological distress.⁵ The impact of housing quality measured in composite has also been demonstrated for mental health. Hwang et al. found that mental health status scores were lower for residents of lower-quality rooming houses,⁶ and Suglia et al. found a higher risk for depression among women experiencing housing disarray and instability (though not housing deterioration).⁷ Harkness and colleagues measured lower mental health service use costs for individuals with mental illness living in higher-quality buildings,⁸ and Wells and Harris demonstrated reduced psychological distress among low-income women after moves to higher-quality housing.⁹ Similarly, both single and composite characteristics of neighborhoods have been associated with health, mental health, and behavioral outcomes.¹⁰ For example, Toomey et al. documented relationships between a higher density of alcohol establishments in urban neighborhoods and rates of four types of violent crime.¹¹ Guite et al. linked specific aspects of the physical and urban environment in London (including noise, overcrowding, absence of green space, community facilities, and fear of crime) with lower mental health status and vitality scores¹² in the general population. Similarly, Matheson and colleagues found higher levels of depression among residents of stressed neighborhoods in a representative sample of over 56,000 adults in 25 census metropolitan areas in Canada.¹³ Moves from distressed to less distressed neighborhoods were followed by improvement in long-term physical health, mental health, and well-being outcomes among low-income adults in a study by Ludwig et al.,¹⁴ and neighborhood-level variables (e.g., perceived low crime levels and good availability of services) predicted reduced psychiatric distress and increased recovery orientation and adaptive functioning among individuals with serious mental illness in a study by Wright and Kloos.¹⁵ Important advances in theory and conceptual understanding of these phenomena have also been made (e.g., O'Campo et al., Fields).^{16,17}

While this accumulation of evidence makes it clear that housing and neighborhood quality are important social and physical ecology variables that may mediate or moderate outcomes in housing intervention research, "objective" housing and/or neighborhood quality is rarely measured or reported in studies of housing interventions for homeless individuals. Studies typically report on the simple

provision of housing without elaborating on quality. If included, housing quality is usually measured either by self-report or only a few observed indicators. This circumstance may be attributed to a lack of standard, reliable, and valid objective housing quality instruments. Keall et al. reviewed the state of the science on housing quality assessment and noted the need for a standard, valid, reliable, and practical way of measuring housing quality to support research, practice, and policy.¹⁸

This article describes the development, pretesting, and piloting of a housing quality instrument for a multisite intervention trial investigating the effectiveness of a housing program for homeless adults with serious mental health and addiction issues in five Canadian cities. The intervention, known as Housing First (HF),¹⁹ involves immediate provision of housing and supports, usually through rent subsidies and in scattered site apartments in the private housing market, which is not conditional on treatment compliance and sobriety. Two thousand one hundred and forty-eight participants in intervention (INT) ($n=1,158$) and treatment-as-usual (TAU) ($n=990$) groups were interviewed every 3 months for up to 24 months, on measures of housing stability, mental illness symptoms, functioning, and quality of life. The trial protocol is provided in detail in Goering et al.²⁰

Documenting the quality of housing in the trial (both as rated subjectively by participants and as assessed by researchers) was considered to be important for examination of its potential relationship with outcomes as well as expected differences between study groups and sites. For example, the cities differed in rental market conditions, so the challenge of finding quality rental units was greater in some sites. We also considered quantifying housing quality to be important for gaining better understanding of the type of housing provided through the publicly funded study intervention and to provide a tool to enable future programs to document housing quality in a uniform way across jurisdictions.

Quality of housing from the participants' perspective ("subjective" housing quality) was also collected in the trial using the Perceived Housing Quality Scale, a set of housing-related satisfaction items selected from two sources.^{21,22} However, we were concerned that ratings collected via self-report might be inaccurate or biased, especially for individuals afraid of losing their housing.

With respect to actual ("objective") housing quality, we initially searched the gray and peer-reviewed literature for a suitable instrument. We identified several tools designed for housing quality inspection purposes, but most were excessively burdensome, and none had the desirable scale properties necessary for a research study. Specifically, items were typically rated as present or absent, with or without nonstandard qualitative comments—which did not provide for standardized, dimensional measurement. Only one housing quality scale developed for research purposes was identified in the peer-reviewed literature. This scale, by Evans et al.,²³ provides for observer ratings in the domains of privacy, cleanliness and clutter, structural quality, and hazards. The scale is lengthy, available in 88- and 53-item versions, and assesses multiple interior rooms separately as well as exterior condition. The scale has been used in urban and rural samples (Michigan and New York, respectively) but primarily for the assessment of housing settings for mothers with children. As such, many of the items were not applicable to adults in sole-living situations in the lower-cost end of the rental housing market, which was our target population and setting. For example, one subscale, "Child Resources," includes items such as "Toys are accessible to the child." In addition, the subscale Cleanliness/Clutter captures occupant housekeeping practices, and our intent was to assess the physical quality of the dwelling itself, not the participants' housekeeping practices. Moreover, some other possible predictors of mental health/wellness of

interest to us were not assessed (e.g., natural light as a variable related to mood). Finally, the scale did not assess aspects of the built environment at the level of the neighborhood. Ultimately, we opted to develop a purpose-built observer-rated housing quality scale for the trial.

METHODS

The instrument was conceptualized as a measure of observed physical quality of the built environment including the housing unit, building, and neighborhood, in the context of recently homeless but now housed adults, living with mental health and/or substance use issues. The pilot study was reviewed and approved as a revision to the main study protocol by academic research ethics boards in the respective pilot jurisdictions. The major steps of development are shown in Fig. 1.

Instrument Development

Item Generation OHQS development was guided by current standards in the literature for the development of scales.^{24–26} The first step involved compilation of a comprehensive list of housing quality characteristics/attributes from the peer-reviewed and gray housing literature. We combined 29 articles from our initial, general literature review on housing quality with 21 from a more focused literature review on housing quality measurement. For the latter, Scholars Portal* was used to search and access articles on combinations of the query keywords: *housing, residential, quality, quality control, satisfaction, inspection, scale, tool, assessment, evaluation, and objective*. No date limiters were used, but the search was restricted to articles in English. From the total yield of 31 abstracts (dating from 1970 to the present), 21 met our relevance criterion, in that they directly discussed methods of evaluating objective housing quality that would apply to dwellings in developed countries. Three additional relevant articles were nominated by members of our study team.

Next, a long list of more than 550 nonunique attributes of housing quality was systematically extracted from the 53 articles and grouped into those descriptive of the unit, the building, and the neighborhood. Additional attributes were extracted from 14 gray literature documents found in a brief internet search (using the same key words listed above) or nominated by colleagues. Attributes representing similar concepts were collapsed by consensus of two raters to form a shorter list of 60 unique attributes (23 for the unit, 17 for the building, and 20 for the neighborhood).

To ensure that the initial bank of housing quality attributes was relevant to our context of adult homelessness and that no important characteristics had been missed, in the next step, we held three focus groups—two with individuals with lived experience of mental health issues and homelessness drawn from the study’s local and national consumer† panels and one with housing service providers drawn from the study’s cross-site housing community of practice. Each group involved 7 to 15

*Scholars Portal is a service of the Ontario Council of University Libraries (OCUL), which provides a stable, integrated search platform to facilitate discovery and access to digital academic literature within multiple disciplines (WLU, 2012). At the time of the search (Spring 2012), the Scholars Portal electronic journal collection included more than 20 million articles from over 8,000 academic journals, published by major distributors and presses (Western Libraries, 2010).

†We recognize that there is no ideal term to refer to all individuals with prior or recent lived experience of homelessness and/or mental illness; in this article, for brevity, we use the term “consumer.”

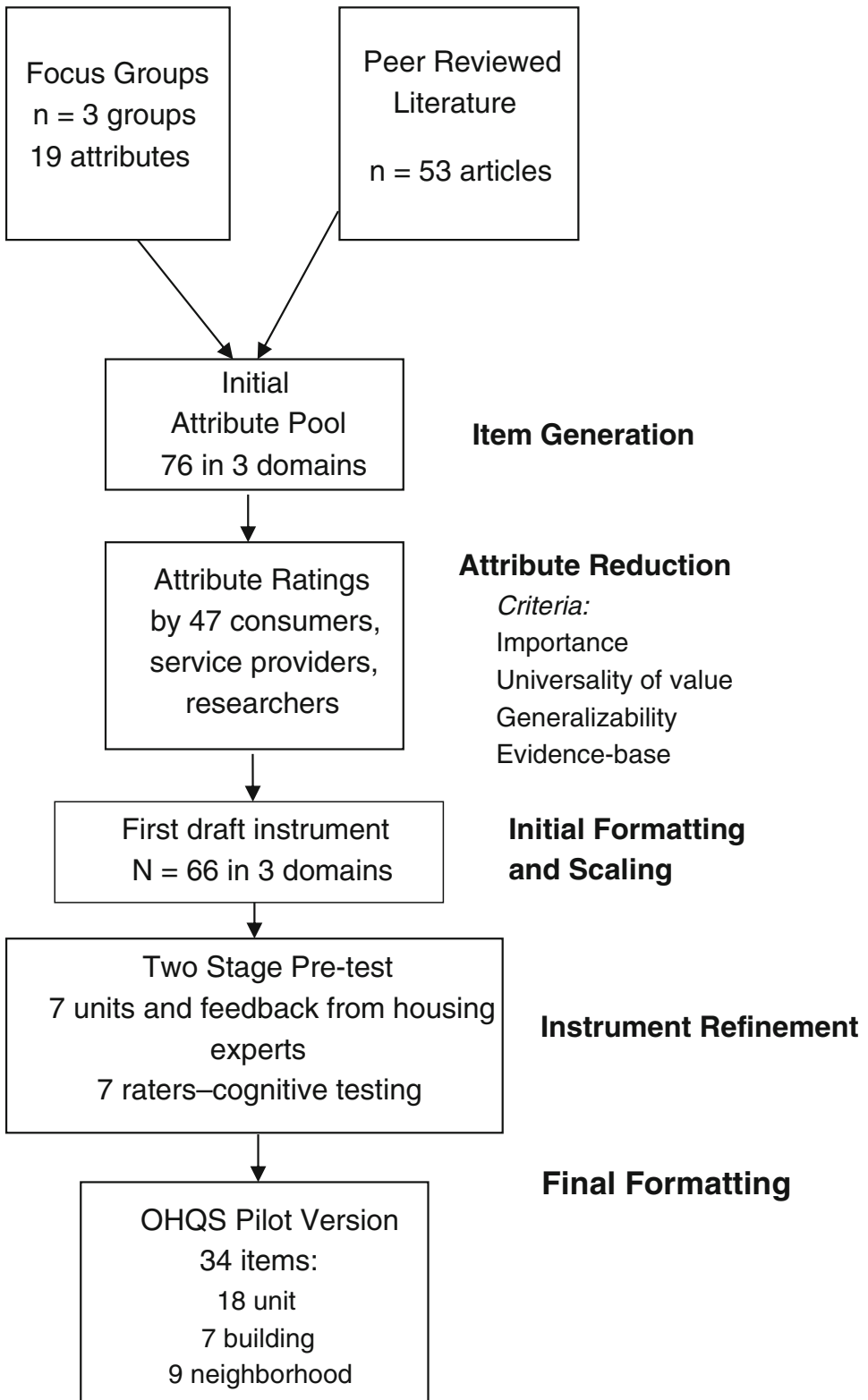


FIGURE 1. Instrument development steps.

participants, and participants without other compensation for their time were given a \$30.00 CAD honorarium. Discussion was guided by questions about best and worst physical features of units, buildings, and neighborhoods, and the most important attributes for mental health and wellness. Notes were content analyzed for attributes of the unit, building, and neighborhood, and attributes not previously identified were added to the master list. For example, the items *lighting*, *noise*, and *available services* were added to the building list. Both the frequency of focus group mention for attributes previously identified as well as the stated importance of particular attributes (especially where general group consensus was noted) was documented to inform the next stage of attribute reduction. Notably, groups consistently considered the unit to be the most important aspect of housing quality, followed by the neighborhood and then by the building, which guided our consideration of the number of final items to include in each category for appropriate overall balance of the final instrument.

Attribute Reduction A set of criteria for selection of a shorter list of “best” attributes from the list of 79 to this stage was developed. They were *perceived importance of the characteristic to overall housing quality*, *generalizability* (i.e., the majority of homes assessed would have the characteristic), *universality of value* (there would be nearly universal agreement about the value of the characteristic (e.g., all agree that the presence of pests is negative)), and *evidence base* (i.e., there was scientific evidence or at least plausibility of an association with health (e.g., natural light and mood)). To assess perceived importance and universality of value, the list was formatted to allow individual written responses and circulated to study stakeholders including consumers, housing and service providers, investigators, research field teams, and site coordinators). Forty-seven respondents rated the attributes as of high, medium, low, or no importance, as well as identified the five most important items in the unit, building, and neighborhood, respectively. Ratings were summarized across respondents to identify the most favored items for each domain. The project committee, through group deliberation, reviewed the list for generalizability and evidence base. In addition, several characteristics were noted by the project committee to be descriptive rather than evaluative so were collected separate from the scale items for general descriptive purposes. These included building and unit type, unit size, tenancy terms (e.g., re: pets, smoking), length of residency, rent payment amounts, and lease structure.

Initial Formatting and Scaling Next, items with rating responses were drafted for each remaining attribute ($n=66$). Response scale anchors were crafted to be as concrete and observable as possible, initially on five-point ordinal scales. Several rounds of project committee discussion and a focused meeting with an experienced housing team in Winnipeg guided refinement of scale anchors. Some items were also identified that would not necessarily be able to be rated at a single time point (e.g., frequent power outages), so for those items, brief questions were developed which could be used to elicit the occupant’s experiences over a period of time (operationalized as the past 3 months).

Pretest The draft questionnaire items were then pretested using cognitive interviewing techniques^{27–29} with seven research assistants (RAs) selected from the field research teams that would ultimately be administering the scale in all five study sites. Cognitive interviewing methods enable systematic assessment of

comprehension and appropriateness of questionnaire instructions, and item wording. Pretesting confirmed that comprehension was good for most terms used in the draft instrument (e.g., weather stripping, sprinkler systems, deadbolt locks), but for a few terms (e.g., escape routes, external focal points), interpretations varied, so more detailed descriptors or definitions were provided and training materials refined. Points of confusion were also noted for extra attention in training. The next revision of the instrument was then pretested in three vacant rental properties and four occupied rental properties known to the research team, followed by further revision. Among other observations, the pretest raters noted that occasionally their rating fell “between” the major categories, so the response scale was revised at this stage to allow for half-point ratings.

Pilot The pilot version of the scale had 18 items on the housing unit itself, seven on the building and nine on the neighborhood. Three neighborhood items were most efficiently populated using external information (Walk Score™,³⁰ distance to green space, and distance to transit), so those were not rated in the pilot, and the neighborhood subscale will not be further discussed in this article. For the 18 housing unit and 7 building items, the response scale ranges from 0.5 to 5 in half-point increments, for a total possible score (across unit and building subscales) ranging from 12.5 to 125 (see example item in Fig. 2).

The objectives of the pilot were to (a) identify any remaining concerns related to face validity and feasibility of scale administration in the field and to (b) establish initial psychometric properties, especially inter-rater reliability.

Training for the pilot consisted of a teleconference presentation outlining the intent of each item and conditions which would represent each scale value, and the assessment of two units at each site by two RAs and a housing expert, who completed the ratings with concurrent discussion and took detailed notes. The housing expert provided coaching on features of housing quality during the training and then provided independent ratings.

For the pilot itself, items were rated independently by two trained RAs and a housing expert in each of 55 units (18 in Toronto, Ontario, and 37 in Winnipeg, Manitoba). At or around the time of the study, Toronto had a population of just over 5.5 million and a 1.6 % vacancy rate, and the average monthly rent for a one-bedroom apartment was \$969 CAD and 5.9 % of homes were reported to be in need of major repairs.^{31,32} Winnipeg had a population of just over 700,000 and

Natural Light	.5	1	1.5	2	2.5	3	3.5	4	4.5	5
	No source of natural light—space is dark if not artificially lit.		Small or few windows and/or no fixtures for appropriate window coverings; e.g., resident has to use unconventional means to cover windows such as tin foil or plastic. Space is dim.			Smaller or few windows relative to space, with basic fixtures for coverings. Space has average brightness.		At least two moderate to large windows and fixtures that would allow for adequate coverings. Space is bright.		Windows large in relation to space in every or nearly every room (except bathroom) incl. windows on opposite or adjacent walls. Space is very bright. Drapery fixtures that permit good quality coverings.

FIGURE 2. Example item from the pilot version.

a .7 % vacancy rate; the average monthly rent for a one-bedroom apartment was \$657 CAD, and 8.4 % of homes were reported to be in need of major repairs.^{31,32}

Units were sampled from a list from the main trial based on information collected in the most recent participant interviews. Included were units occupied by the participant for at least 3 months (this was necessary because a 3-month recall period was used for the interview-based items). Excluded were the units of participants who could not be reached after three attempts or those of participants who would require an interpreter for the interview component. In a very small number of cases, RAs had safety concerns about visiting a participant due to some history of threatening behavior and site investigators made case-by-case decisions about the inclusion of these individuals. Three TAU participants were sampled for every two INT participants because it was expected that more variability would be seen in TAU housing circumstances, and we wanted to ensure a sufficient sample to capture that variability. Participants were contacted by phone, explained the purpose and process of the visit, and assured that participation was voluntary. They were told that the researchers were not interested in how the place was kept, but rather the quality of the structure itself (doors, window, appliances), that RAs may need to check on specific aspects of the unit (e.g., look under sinks), that they would also be asked some questions about their place during the visit, and that they would be given \$20 CAD for their participation.

During the visits, one RA recorded the participant's responses verbatim, and the other made brief notes as needed to inform ratings. RAs could discuss characteristics of the space during the visit but not reveal or compare their ratings. Given that no single description can represent all variation in housing characteristics, RAs were instructed to choose the category representing "best fit" rather than "perfect fit" and to note items that were challenging to rate for any reason. During the visit, the items were rated on paper forms, and the data were subsequently entered to the main study database. Given the potential for transfer of bedbugs from infestations in some units in each site, a standard prevention protocol was used, and no such incidents occurred. Visits took on average 45 min for assessment of all three components.

Data Analysis

Item, subscale (unit and building), and total scale score (for unit and building items) distributions were first inspected for normality, floor, and ceiling effects. Forty-two units were rated by three observers, 9 by two observers, and 4 by a single observer. For the purposes of calculating descriptive statistics, observations were weighted by the inverse of the number of ratings given. This was done in order to give each housing unit equal weight in this part of the analysis. Totals were rescaled to take missing items into account. A maximum of two missing values were permitted for the unit subscale and one for the building subscale. Totals were not calculated for eight records (5.4 % of the total) with three or more missing items.

Summary statistics, including minimum, maximum, mean, and standard deviation, were produced for each item and subscale. Percent agreement and intraclass correlation coefficients (ICC (1, 1)) were used to examine inter-rater reliability (IRR) of nonexpert raters. The ICCs reported reflect estimates of scale reliability with a single rater. Item-total correlations, subscale-to-total correlations, and Cronbach's alpha for the unit and building subscales together were calculated to assess internal consistency. Finally, equality of item and subscale score distributions

were examined by site, study group (intervention or treatment as usual), individual rater, and type of rater (lay or expert) using Kolmogorov–Smirnov statistics. To test differences between sites in agreement, for ICCs, we used R to Z transformations, and to test differences in housing quality between study groups and sites, we used Student's t tests.

RESULTS

Item Analysis

Results are presented in Table 1 for each item. Five items (*utilities-power*, *artificial light*, *utilities-heating/cooling*, *utilities-water*, and *utilities-plumbing*) from the unit subscale and three items (*security/safety*, *staff in building*, and *access/visitability*) from the building subscale displayed ceiling or floor effects of a nature that adjustments to the scale anchors were required. Percent agreement ranged from 89 to 100 % within one response scale value and 67 to 91 % within a half response scale value (data not shown). Four items had two-rater ICCs below 0.74 (*artificial*

TABLE 1 Item characteristics

Item	Mean (SD)	%A (1) ^a	ICC (2) ^b (95 % CI)	ITC ^c
Unit				
1. Safety/security	3.9 (1.0)	96	0.74 (0.58–0.85)	0.68
2. Natural light	3.9 (0.90)	96	0.73 (0.57–0.84)	0.64
3. Artificial light	4.1 (0.80)	90	0.58 (0.36–0.74)	0.47
4. Utilities-power	4.3 (0.90)	98	0.78 (0.63–0.87)	0.70
5. Indoor air/ventilation	3.8 (0.90)	96	0.77 (0.63–0.87)	0.68
6. Utilities-heating/cooling	4.1 (0.90)	96	0.83 (0.72–0.90)	0.45
7. Utilities-water	4.5 (0.70)	98	0.82 (0.71–0.90)	0.56
8. Utilities-plumbing	4.3 (1.0)	98	0.86 (0.76–0.92)	0.66
9. Bathroom facilities	3.4 (1.4)	92	0.92 (0.86–0.95)	0.64
10. Structural condition	3.8 (1.1)	100	0.91 (0.85–0.95)	0.83
11. Kitchen/food prep area	3.5 (1.4)	96	0.85 (0.75–0.92)	0.61
12. Kitchen appliances	3.6 (1.2)	98	0.90 (0.82–0.94)	0.60
13. Bedroom/sleeping space	3.8 (1.1)	94	0.74 (0.59–0.85)	0.41
14. Noise	3.6 (1.1)	92	0.68 (0.49–0.80)	0.46
15. Pests	3.9 (1.3)	96	0.94 (0.89–0.96)	0.46
16. Storage space	3.5 (1.3)	92	0.87 (0.78–0.92)	0.59
17. Overall design	3.1 (1.1)	89	0.76 (0.61–0.86)	0.62
18. Laundry	3.0 (1.2)	98	0.93 (0.88–0.96)	0.66
Building				
19. Security/safety	4.0 (1.1)	98	0.90 (0.82–0.94)	0.74
20. Staff in building	4.1 (1.1)	98	0.87 (0.78–0.93)	0.47
21. Access/visitability	2.4 (1.6)	96	0.94 (0.89–0.96)	0.50
22. Inside condition (common areas)	3.7 (1.2)	98	0.83 (0.72–0.90)	0.85
23. Outside condition (property and building)	3.7 (1.0)	100	0.89 (0.81–0.94)	0.67
24. Garbage facilities	3.7 (1.2)	100	0.68 (0.49–0.81)	0.51
25. Access to nature (on property)	2.8 (1.3)	96	0.92 (0.86–0.95)	0.54

^aPercent agreement within one response value

^bIntraclass correlation coefficient for the two nonexpert raters

^cItem-to-total-score correlation coefficient

light; bedroom/sleeping space; noise and garbage) and these were targeted for further refinement of response descriptors. While some higher ICCs may be attributable in part to the lack of spread in the distributions, this did not appear to be the dominant explanation except for access/visitability. Correlations between individual items and the full scale score were in the desired range (not so high that the item provides no additional information to the scale and not so low that the item is inconsistent with the overall construct).

Subscales and Total Score

Subscale and total score distributions had reasonably good spread and moderate skew. The mean total score overall was 92.48 (SD 16.56; range 50.69 to 112.68). Internal consistency was good in all cases (0.80 for the building; 0.90 for the unit; 0.92 for the total across the two subscales). Inter-rater reliability was also very high for the two subscales and total score, with respective estimates of 0.87 (unit), 0.85 (building), and 0.93 (total score).

Distributions of ratings did not differ significantly across either individual raters or between expert and nonexpert raters. Reliability did vary by site for some items—significant differences were found for 7 of the 25 items at the 0.05 level (*natural light, indoor air/ventilation, kitchen food prep area, pests, storage, overall design, and staff in building*), and one at the 0.01 significance level (indoor air/ventilation). A statistically significant difference between sites in reliability of the building subscale was found, but not the unit or total (Table 2).

Statistically significant differences in housing quality for the unit and total score (but not the building) were found between INT and TAU groups, with total quality lower in TAU and much greater variability in TAU (Table 3; Fig. 3). Differences in housing quality by site were in the expected direction, but were not significant with this sample size.

DISCUSSION

We developed and piloted a standard scale to be used by trained, nonexpert RAs to measure the physical quality of housing (units, buildings, and neighborhoods) occupied by recently homeless adults living with mental health issues. Once the scale was constructed and pretested, we found training and administration to be feasible and straightforward. Despite some concern about the ability to capture multiple and often complex attributes of housing structure and facilities, we found good reliability at the item, subscale, and total score levels of the new instrument. For example, one item which was predicted to have low IRR was *design* (RAs felt it was very subjective), yet it had a very acceptable ICC at 0.76.

We have initially attributed poorer distributions and/or lower reliability to item construction and have revised items accordingly. However, it may be that the

TABLE 2 Reliability by site for subscales and total score

	WINN Rho (95 % CI) (n=37)	TO Rho (95 % CI) (n=18)	Difference Z	p
Unit	0.84 (0.70–0.92)	0.93 (0.82–0.97)	1.28	0.20
Building	0.89 (0.79–0.94)	0.98 (0.94–0.99)	2.55	0.01
Total	0.89 (0.79–0.95)	0.97 (0.92–0.99)	1.91	0.06

TABLE 3 Total housing quality scores (analyses based on mean ratings across raters per residence) by site and study group

	INT mean (SD) (n=18)	TAU mean (SD) (n=37)	t	p
Unit	76.68 (12.36)	63.85 (12.36)	-5.69 ^a	0.000
Building	25.52 (3.55)	23.90 (6.35)	-1.22	0.229
Total	102.21 (5.96)	87.75 (18.0)	-4.41	0.000
	WINN mean (SD) (n=37)	TO mean (SD) (n=18)		
Unit	67.21 (12.15)	69.77 (11.87)	0.739 ^b	0.463
Building	23.42 (5.51)	26.52 (5.35)	-1.97	0.054
Total	90.63 (16.66)	96.28 (16.13)	-1.19	0.238

^aEqual variances not assumed

^bEqual variances assumed

distributions reflect an underlying variable that is not ordinal. For example, very few units were found to have any access/visitation features, yet a small number had most or all *access/visitation* features. Thus, this variable, in most residences, may be a “present or absent” phenomenon. We will examine the degree to which item revisions have alleviated these patterns in a much larger sample in the main study.

Most item-total correlations found seemed plausible. For example, a very high quality unit and building could nevertheless have a high level of noise from an external source, and therefore a low correlation between the *noise* item and the total score. Also, it would be reasonable to predict that the *structural condition* item would correlate very highly with the total score since it would reflect the age, condition, and/or maintenance level of the residence across several functional areas.

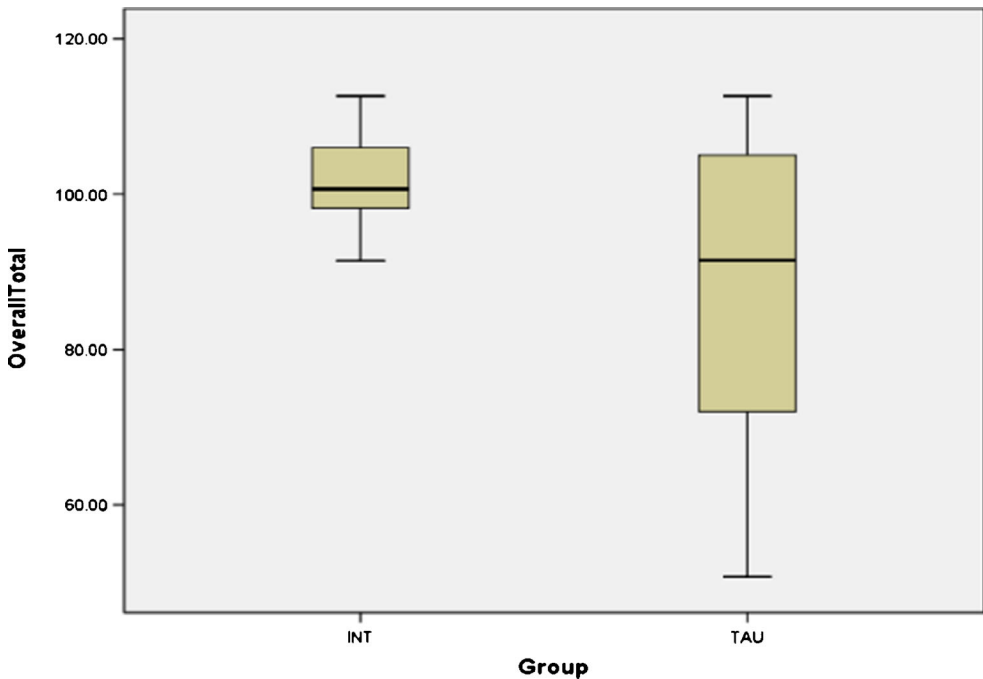


FIGURE 3. Total housing quality scores by study group.

However, for some items (e.g., bedroom/sleeping space), the reason for a low item-total correlation (ITC) was less easy to explain.

Differences in item-level IRR between sites were few and generally of low magnitude, but there was a significant difference between sites for the building subscale. To improve agreement further for the main study, in addition to item modifications, we developed a more intensive 2-day training program that included interactive classroom instruction using photographs representative of rating levels for each item, and more in-depth field practice followed by debriefing.

The differences in measured housing quality by study group were significant and in the expected direction, supporting the sensitivity of the scale for its intended purpose. The scale worked very well in the pilot for the most common type of building in which study participants were housed in the larger study—i.e., multiunit apartment buildings or residences. There were a few problems with applicability of items for less common building types such as single family homes. It will be possible to examine these issues in the larger sample in the main study.

Psychometric findings are context specific, and therefore, our findings may not generalize to other resident populations, other geographic settings and climates, and especially housing outside the lower-cost end of the market. However, given that there are so few instruments available for research on the health outcomes of the physical features of housing that can produce dimensional scale scores for research purposes, research on broader applications and/or adaptations may be warranted.

As housing design and construction standards change, relevancy of specific items may also decline. For example, our assumption that standard size appliances are generally of better quality than nonstandard sizes may be increasingly false given the increasing availability of good-quality smaller appliances for small spaces.

The OHQS was revised after the pilot. Modifications included item revisions based on the psychometric findings and reordering based on specific feedback from the pilot RAs about the flow of the visits. No item performed so poorly as to warrant dropping it completely; therefore, all items were retained for use in the main study. Administration began in October 2011 in a sample of 560 participant residences in the five cities of the main trial (Vancouver, Winnipeg, Toronto, Montreal, and Moncton, Canada).

Having an instrument that can quantify housing quality may be important, not only for accurate measurement in housing and health-related research, but also for policy and practice. As communities move toward more HF approaches, such a tool may assist programs with decisions about property inclusion and guiding client choices. For example, it may be useful in the future as part of HF program fidelity assessments as they are implemented in very different housing contexts. At the policy level, it can help ensure accountability for a standard of housing quality in programs funded with public dollars. The final version of the instrument and training materials are expected to be available in 2014.

CONCLUSIONS

Initial, small sample results indicate that the physical quality of housing can be reliably rated by trained but nonexpert raters using the OHQS. The tool has potential for improved measurement in similar housing-related health research, may assist housing programs in practical ways, and may be important for documenting the quality of housing that is provided by publicly funded housing programs.

ACKNOWLEDGMENTS

Thanks to K. Mason, C. Kelly, and C. Issak (for assisting with ethics submissions, recruitment, and supervision of local field teams); T. Bischoff for assistance with attribute reduction; A. Ladd, D. Powell, and R. Galston for site assessments; F. Weinstock and L. Stewart (who served as our housing experts); and D. Streiner (for comments on the analysis).

We also thank Jayne Barker (2008–2011), Ph.D., and Cameron Keller (2011–present), Mental Health Commission of Canada At Home/Chez Soi National Project Leads, the National Research Team, the five site research teams, the site coordinators, and the numerous service and housing providers, as well as persons with lived experience, who have contributed to this project and the research. This research has been made possible through a financial contribution from Health Canada. The views expressed herein solely represent the authors.

REFERENCES

1. Krieger J, Higgins DL. Housing and health: time again for public health action. *Am J Public Health*. 2002; 92(5): 758–768.
2. Osman LM, Ayres JG, Garden C, et al. Home warmth and health status of COPD patients. *Eur J Public Health*. 2008; 18(4): 399–405.
3. Hwang SW, Wilkins R, Tjepkema M, et al. Mortality among residents of shelters, rooming houses, and hotels in Canada: 11 year follow-up study. *BMJ*. 2009; 339: b4036. doi:10.1136/bmj.b4036.
4. Jacobs DE, Brown MJ, Baeder A, et al. A systematic review of housing interventions and health: introduction, methods and summary findings. *J Public Health Manag Pract*. 2010; 16(5): S5–S10.
5. Evans GW. The built environment and mental health. *J Urban Health*. 2003; 80(4): 536–555.
6. Hwang SW, Martin RE, Tolomiczenko GS, et al. The relationship between housing conditions and health status of rooming house residents in Toronto. *Can J Public Health*. 2003; 94(6): 436–440.
7. Suglia SF, Duarte CS, Sandel MT. Housing quality, housing instability and maternal mental health. *J Urban Health*. 2011; 88(6): 1105–1116.
8. Harkness J, Newman SJ, Salkever D. The cost-effectiveness of independent housing for the chronically mentally ill: do housing and neighborhood features matter? *Health Serv Res*. 2004; 39(5): 1341–1360.
9. Wells NM, Harris JD. Housing quality, psychological distress, and the mediating role of social withdrawal: a longitudinal study of low-income women. *J Environ Psychol*. 2007; 27: 69–78.
10. Lindberg RA, Shenassa ED, Acevedo-Garcia D, et al. Housing interventions at the neighborhood level and health: a review of the evidence. *J Public Health Manag Pract*. 2010; 16(5): S44–S52.
11. Toomey TL, Erickson DJ, Carlin BP, et al. The association between density of alcohol establishments and violent crime within urban neighborhoods. *Alcohol Clin Exp Res*. 2012. doi:10.1111/j.1530-0277.2012.01753.x.
12. Guite HF, Clark C, Ackrill G. The impact of the physical and urban environment on mental well-being. *Public Health*. 2006; 120(12): 1117–1126.
13. Matheson FI, Moineddin R, Dunn JR, et al. Urban neighborhoods, chronic stress, gender and depression. *Soc Sci Med*. 2006; 63: 2604–2616.
14. Ludwig J, Duncan GJ, Genetian LA, et al. Neighborhood effects on the long-term well-being of low-income adults. *Sci*. 2012; 337: 1505.

15. Wright PA, Kloos B. Housing environment and mental health outcomes: a levels of analysis perspective. *J Environ Psychol.* 2007; 27(1): 79–89.
16. O'Campo P, Salmon C, Burke J. Neighbourhoods and mental well-being: what are the pathways? *Health Place.* 2009; 15: 56–68.
17. Fields D. Emotional refuge? Dynamics of place and belonging among formerly homeless individuals with mental illness. *Emot Space Soc.* 2010. doi:10.1016/j.emospa.2010.09.002.
18. Keall M, Baker MG, Howden-Chapman P, et al. Assessing housing quality and its impact on health, safety and sustainability. *J Epidemiol Community Health.* 2010; 64: 765–771.
19. Tsemberis S, Gulcur L, Nakae M. Housing First, consumer choice, and harm reduction for homeless individuals with a dual diagnosis. *AJPH.* 2004; 94(4): 651–656. doi:10.2105/AJPH.94.4.651.
20. Goering PN, Streiner DL, Adair C, et al. The At Home/Chez Soi trial protocol: a pragmatic, multi-site, randomized controlled trial of a Housing First intervention for homeless individuals with mental illness in five Canadian cities. *BMJ Open.* 2011; 1: e000323. doi:10.1136/bmjopen-2011-00323.
21. Tsemberis S, Rogers ES, Rodis E, et al. Housing satisfaction for persons with psychiatric disabilities. *J Community Psychol.* 2003; 31(6): 581–590.
22. Toro PA, Bellavia CW, Daeschler CV, et al. Evaluating an intervention for homeless persons: results of a field experiment. *J Consul Clin Psychol.* 1997; 65(3): 476–484.
23. Evans GW, Wells NM, Chan HYE, Saltzman H. Housing quality and mental health. *J Consul Clin Psychol.* 2000; 68(3): 526–530.
24. US DHHS, FDA, CDER, CBER, CDRH. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. December 2009.
25. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess.* 1995; 7(3): 309–319.
26. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use.* 4th ed. Oxford: Oxford University Press; 2008.
27. McColl E, Meadows K, Barofsky I. Cognitive aspects of survey methodology and quality of life assessment. *Qual Life Res.* 2003; 12: 217–218.
28. Willis GB. *Cognitive interviewing: a tool for improving questionnaire design.* Thousand Oaks, CA: Sage; 2005.
29. Collins D. Pretesting survey instruments: an overview of cognitive methods. *Qual Life Res.* 2003; 12: 229–238.
30. Carr LJ, Dunsiger SI, Marcus BH. Validation of Walk Score for estimating access to walkable amenities. *Br J Sports Med.* 2010. doi:10.1136/bjism.2009.069609.
31. Canadian Mortgage and Housing Corporation. Rental market report Canada highlights, Spring 2011.
32. Statistics Canada. Census Canada 2006 <http://www12.statcan.gc.ca/census-recensement/2006/dp-pd/prof/92-591/>. Accessed 20 July 2013.