Singapore Med J 2022; 63(2): 93-96
https://doi.org/10.11622/smedj.2020101

# Inter-rater reliability of a composite health promotion scoring system developed in Singapore

*Manimegalai Kailasam*[1], MBBS, MPH, *Priyanka Vankayalapati*[1], MBBS, MPH, *Yin Maw Hsann*[1], MBBS, MMed, *Kok Soong Yang*[1], MBBS, MMed

**INTRODUCTION** In view of the important role of the environment in improving population health, implementation of health promotion programmes is recommended in living and working environments. Assessing the prevalence of such community health-promoting practices is important to identify gaps and make continuous and tangible improvements to health-promoting environments. We aimed to evaluate the inter-rater reliability of a composite scorecard used to assess the prevalence of community health-promoting practices in Singapore.
**METHODS** Inter-rater reliability for the use of the composite health promotion scorecards was evaluated in eight residential zones in the western region of Singapore. The assessment involved three raters, and each zone was evaluated by two raters. Health-promoting practices in residential zones were assessed based on 44 measurable elements under five domains – community support and resources, healthy behaviours, chronic conditions, mental health and common medical emergencies – in the composite scorecard using weighted kappa. The strength of agreement was determined based on Landis and Koch's classification method.
**RESULTS** A high degree of agreement (almost perfect-to-perfect) was observed between both raters for the measurable elements from most domains and subdomains. An exception was observed for the community support and resources domain, where there was a lower degree of agreement between the raters for a few elements.
**CONCLUSION** The composite scorecard demonstrated a high degree of reliability and yielded similar scores for the same residential zone, even when used by different raters.

*Keywords: composite, health promotion, inter-rater reliability, score*

## INTRODUCTION

A health-promoting environment is made up of health-promoting programmes, policies, practices and aspects of the built-up environment.[1,2] Its role in preventing chronic diseases and improving the health of the population has been supported by substantial evidence in the existing literature.[3-5] For example, health promotion programmes implemented in workplaces are associated with positive outcomes, such as reduced absenteeism and increased work ability in workers.[6,7] In addition, it has demonstrated positive financial outcomes, such as low medical and absenteeism costs, and substantial cost savings.[8,9] In view of these benefits, implementation of health promotion programmes ought to be recommended in living and working environments.

Composite scores could be used as a standardised method to evaluate health promotion programmes and ensure that they are comprehensive and evidence based. The CDC (Centers for Disease Control and Prevention) Worksite Health ScoreCard (HSC) is used to assess employers' health promotion practices in workplaces in the United States.[10] In Singapore, we developed a similar composite health promotion scorecard after consulting experts from the Saw Swee Hock School of Public Health, National University of Singapore, and the Health Promotion Board to assess the prevalence of health-promoting practices in residential zones. The scorecard consisted of 44 measurable elements that were associated with positive lifestyle and health behaviour changes identified from a comprehensive review of the existing literature. These measurable elements were grouped under five domains: community support and resources; healthy behaviour; chronic conditions; mental health; and common medical emergencies. Each element was assigned a weightage based on the strength of evidence and its impact. Some examples of the measurable elements under each of the five domains of the composite scorecard are listed individually in Table I.[11]

Being an aggregate measure of multiple performance indicators using a predetermined weighting methodology,[12] composite scores are subject to measurement errors. Similar to new instruments of health measurement scales, newly developed composite scores ought to be validated before their formal use. The HSC is a validated tool – its face validity and inter-rater reliability were investigated[13] before it was used to evaluate health promotion programmes in workplaces. Similarly, the composite health promotion scorecard developed in Singapore should be validated before its usage in residential communities. As its face validity had been determined by experts during the development phase,[11] we explored the reliability issue due to the involvement of different assessors when using the composite scorecard. We aimed to evaluate the inter-rater reliability of this composite scorecard for assessing the health-promoting environment in residential communities.

[1]Department of Epidemiology, Ng Teng Fong General Hospital, Singapore
**Correspondence:** Dr Manimegalai Kailasam, Senior Epidemiologist, Epidemiology, Ng Teng Fong General Hospital, 1 Jurong East Street 21, Singapore 609606. Kailasam_Manimegalai@nuhs.edu.sg

**Table I. Examples of measurable elements grouped under five domains of the composite health promotion scorecard.**

| Scorecard domain | Measurable element |
|---|---|
| Community support and resources | Health promotion committee |
| | Publicity of health promotion programmes |
| | Literacy/culture appropriate health promotion programmes |
| Healthy behaviour | Adequate exercise facilities |
| | Healthier choices |
| | Tobacco cessation programmes |
| | Weight management programmes |
| Chronic conditions | Free/subsidised health screening to detect chronic conditions |
| | Talks and training to caregivers of the elderly |
| | Self-management programmes for chronic conditions |
| Mental health | Support system to provide tangible assistance |
| | Support system to provide social and emotional support |
| | Age-appropriate life skills training programmes |
| Common medical emergencies | AED equipped in community centre |
| | AEDs are routinely maintained and tested |
| | Access to training on CPR/AED for residents |

AED: automated external defibrillator; CPR: cardiopulmonary resuscitation

## METHODS

The composite health promotion scorecard was developed through a number of processes, such as a review of health promotion literature to identify elements, including interventions, pertinent to changing individual lifestyle and health behaviour. This list was further reviewed to select elements that were relevant to a residential community. Additionally, national health promotion guidelines were incorporated to suit the local context.[11]

The inter-rater reliability of the composite scorecard was assessed in eight residential zones in the western region of Singapore. These zones were community subdivisions of two different electoral constituencies. Guidelines were developed to assist in and bring consistency to its scoring. Each measurable element was scored as ‘fully met’, ‘partially met’ or ‘not met’ based on whether the zone met the criteria specified in the guidelines. The score for each element under a domain was then added to obtain the domain score for the residential zone. The overall score for the residential zone was a total of the various domain scores.[11]

A total of three raters were involved in the assessment. For each zone, two raters took turns to conduct site visits and interview grassroots leaders and committee members of the zone. Discussions with grassroots leaders and observations made during site visits were documented separately by these raters. The raters then used the composite scorecard and scoring guidelines to appraise the health-promoting environments and score the residential zones independently.

The scoring of each residential zone was done by two raters. These raters were medical graduates with a Master’s degree in public health and a minimum of two years’ working experience in hospital epidemiology.[11] Raters were trained by an experienced senior public health consultant who developed the scoring methodology based on international hospital quality accreditation standards.

The training session spanned a total of three hours and included the following: (a) general overview and discussion on the scoring methodology/guidelines for the scoring system; (b) item-by-item review of the measurable elements of the composite health promotion scoring system to train staff in the intent, assessment process and scoring of individual elements; (c) completion of a practice assessment on a hypothetical scenario; and (d) group discussion of the practice assessments to resolve any areas of confusion. Any queries regarding the measurable elements and scoring guidelines were clarified by the trainer.

Inter-rater reliability was assessed for every measurable element scored by both raters in the composite scorecard using weighted kappa, with a range of 0 (no agreement) to 1 (perfect agreement). Weighted kappa was used instead of Cohen’s kappa (unweighted kappa), as the former takes the ordered nature of the weightage values into account. In contrast, Cohen’s kappa assumes the degree of disagreement between two raters to be the same for all pairs of weightage values. For example, using Cohen’s kappa, [1,2] and [1,3] weighted score pairs would be viewed to have the same degree of disagreement. However, weighted kappa would take into account that the degree of disagreement would be larger for the [1,3] weighted score pair than the [1,2] weighted score pair. Hence, weighted kappa is able to perceive a difference in the degree of disagreement between the raters.

The strength of agreement between Raters 1 and 2 was determined using Landis and Koch’s classification method: slight (≤ 0.20); fair (range 0.21–0.40); moderate (range 0.41–0.60); substantial (range 0.61–0.80); and almost perfect (range 0.81–1.00).[14]

## RESULTS

The classification of these elements is summarised by chapter and subchapter in Table II. The composite score was found to be the same across all zones and between the two raters for some measurable elements. In this unique situation, a weighted kappa value could not be calculated and classified into any of the categories that were proposed by Landis and Koch. Therefore, a new category, ‘perfect’, was added to this study to account for this scenario.

The weighted kappa values for the measurable elements were concentrated in the ‘almost perfect’ and ‘perfect’ categories for most chapters and subchapters, implying a high degree of agreement between Raters 1 and 2 in our study. A dissimilar trend was observed for the domain of community support and resources where, despite most of the measurable elements being clustered in the almost perfect and perfect categories, a couple of elements were classified into the slight and substantial categories. Therefore, Raters 1 and 2 had a varying degree of agreement for some elements of this domain.

**Table II. Strength of agreement by domain based on Landis and Koch's classification method.**

| Composite scorecard domain | No. of elements | | | | | | |
|---|---|---|---|---|---|---|---|
| | Total (n = 44) | Strength of agreement | | | | | |
| | | Slight | Fair | Moderate | Substantial | Almost perfect | Perfect* |
| **Community support and resources** | 11 | 1 | 0 | 0 | 1 | 6 | 3 |
| **Healthy behaviour** | | | | | | | |
| Physical activity | 5 | 0 | 0 | 0 | 0 | 3 | 2 |
| Healthy eating | 4 | 0 | 0 | 0 | 0 | 4 | 0 |
| Smoking prevention | 3 | 0 | 0 | 0 | 0 | 1 | 2 |
| Weight management | 3 | 0 | 0 | 0 | 0 | 3 | 0 |
| **Chronic conditions** | 7 | 0 | 0 | 0 | 0 | 7 | 0 |
| **Mental health** | 5 | 0 | 0 | 0 | 0 | 2 | 3 |
| **Common medical emergencies** | 6 | 0 | 0 | 0 | 0 | 1 | 5 |

Strength of agreement was determined using Landis and Koch's classification method: slight (≤ 0.20), fair (range 0.21–0.40), moderate (range 0.41–0.60), substantial (range 0.61–0.80), almost perfect (range 0.81–1.00) and perfect (complete agreement).

## DISCUSSION

It is important to ensure minimal measurement errors when multiple assessors are associated with the use of composite health promotion scorecards, and accordingly, this study aimed to examine its inter-rater reliability. We found a high degree of agreement for most measurable elements from the domains and subdomains based on Landis and Koch's classification method. However, a lower degree of agreement was found for a couple of elements from the community support and resources domain. The disagreement could probably be attributed to some differences in the interpretation of interview statements provided by the residential zones.

The use of weighted kappa, as an appropriate inter-rater reliability measure, was advantageous for our study. Weighted kappa was preferred over unweighted kappa, as it would take the ordinal nature of the three-level rating system and the relative differences between levels into consideration,[15] thereby improving the accuracy of the study's inter-rater reliability. Furthermore, the kappa statistic is a more superior measure than the conventional percent agreement owing to its ability to account for chance agreement.[16] Both raters used the scoring guidelines of the composite scorecard, thereby minimising potential variability associated with subjective scoring preferences.

The main limitation was the small number of raters and residential zones being assessed in the study, thereby affecting the accuracy of the inter-rater reliability measured. As this was a pilot study, it was conducted only among a limited number of constituency zones. It could also be argued that the high degree of agreement was on account of the background of the raters, who were epidemiologists with formal training in public health. However, scoring guidelines were created in a manner that would be simple for non-professionals to use. Furthermore, the feasibility of the score and the face validity of the scoring guidelines were evaluated by grassroots leaders in the community. Nonetheless, it is still pertinent that the current raters were not the eventual intended users of the composite scorecard, and this has been identified as an area for improvement in future studies.

Owing to the nature of implementation on the ground, there was a lack of complete independence in the assessment of the residential zones by the two raters. For example, part of the assessment for the scoring involved meeting with a group of committee members, including chairpersons of the zone, to obtain information about the zone's programmes and practices. However, this could not be conducted separately for the raters, as it was not practically feasible to organise multiple sessions. Nevertheless, as the raters referenced the scoring guidelines and scored independently of each other, this is unlikely to have overtly influenced the end scores given to the zones.

In conclusion, the composite health promotion scorecard yielded similar scores for the same residential zone, even when used repeatedly by different users. Further exploration could involve non-professional raters in assessments of residential zones.

## ACKNOWLEDGEMENTS

We would like to thank Mr Johnny Wong for his help with the statistical analysis.

## REFERENCES

1. Burton J; World Health Organization. WHO Healthy Workplace Framework and Model: background and supporting literature and practices. Available at: http://www.who.int/occupational_health/healthy_workplace_framework.pdf. Accessed March 13, 2018.
2. Department of Human Services, State Government of Victoria, Australia. Environments for Health: Promoting Health and Wellbeing through Built, Social, Economic and Natural Environments. Municipal Public Health Planning Framework. Available at: https://www.healthyplaces.org.au/userfiles/file/Environments%20for%20Health%20Victoria.pdf. Accessed March 6, 2018.
3. Malambo P, Kengne AP, De Villiers A, Lambert EV, Puoane T. Built environment, selected risk factors and major cardiovascular disease outcomes: a systematic review. PLoS One 2016; 11:e0166846.
4. Adam A, Jensen JD. What is the effectiveness of obesity related interventions at retail grocery stores and supermarkets? A systematic review. BMC Public Health 2016; 16:1247.
5. Bird E, Ige J, Burgess-Allen J, Pinto A, Pilkington P; Public Health and Wellbeing Research Group. Healthy people healthy places evidence tool: evidence and practical linkage for design, planning and health. Available at: http://eprints.uwe.ac.uk/31390. Accessed March 6, 2018.
6. Kuoppala J, Lamminpää A, Husman P. Work health promotion, job well-being, and sickness absences--a systematic review and meta-analysis. J Occup Environ Med 2008; 50:1216-27.

7. Pelletier KR. A review and analysis of the clinical and cost-effectiveness studies of comprehensive health promotion and disease management programs at the worksite: update VIII 2008 to 2010. J Occup Environ Med 2011; 53:1310-31.
8. Baicker K, Cutler D, Song Z. Workplace wellness programs can generate savings. Health Aff (Millwood) 2010; 29:304-11.
9. Chapman LS; American Journal of Health Promotion Inc. Meta-evaluation of worksite health promotion economic return studies: 2005 update. Am J Health Promot 2005; 19:1-11.
10. US Centers for Disease Control and Prevention. The CDC Worksite Health ScoreCard: an assessment tool for employers to prevent heart disease, stroke, and related health conditions. Updated January 2014. Available at: https://www.cdc.gov/dhdsp/pubs/docs/hsc_manual.pdf. Accessed April 4, 2017.
11. Kailasam M, Hsann YM, Vankayalapati P, Yang KS. Prevalence of community health-promoting practices in Singapore. Health Promot Int 2019; 34:447-53.
12. Austin JM, D'Andrea G, Birkmeyer JD, et al. Safety in numbers: the development of Leapfrog's composite patient safety score for U.S. hospitals. J Patient Saf 2014; 10:64-71.
13. Roemer EC, Kent KB, Samoly DK, et al. Reliability and validity testing of the CDC Worksite Health ScoreCard: an assessment tool to help employers prevent heart disease, stroke, and related health conditions. J Occup Environ Med 2013; 55:520-6.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.
15. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005; 37:360-3.
16. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22:276-82.