Introduction

Blue spaces are outdoor environments - formed naturally or by people - that prominently feature water (e.g. rivers, lakes, ponds, canals, fountains, coastal margins, Grellier et al. 2017, p.3). As well as their physical properties (size, shape, clarity, etc.) they also embody a range of socio-cultural and spiritual values (Völker and Kistemann, 2011, 2013, 2015).

Historically, coastal or inland water bodies have been attractive residential locations supplying many provisioning, regulatory and cultural ecosystem services and benefits (Solomon, 2010). Many cities lie on water bodies (e.g. sea, river, lakes) and this has shaped their character (Grellier et al, 2017; Solomon, 2010; Strauss, 2002). Post-industrial cities have witnessed a decline of waterfront industries such as ports (Hoyle, 2002; Hoyle and Pinder, 1992) and consequently much abandoned waterfront land has accumulated (Hoyle and Pinder, 1981;1992). Recently, many post-industrial waterfront cities have transformed these abandoned waterfronts into thriving urban public spaces (Breen and Rigby, 1994; Cary-Elwes,1996; Feldman M. 1999).

The relationship between water and human health has been explored for centuries (Foley and Kistermann, 2015). The therapeutic use of water goes back to Roman times (e.g. Bath in England) and in the 18th century, sea bathing became popular in parts of Europe due to the various claimed prophylactic properties of seawater (Walton, 1983; Corbin, 1994). Blue spaces are increasingly recognised as a source of many benefits but also risks, while they face many environmental, social, and economic challenges (Grellier et al, 2017). Moreover, for environmental justice, accessibility to clean and healthy urban blue settings is needed to ensure that the restorative qualities for recreation and relaxation and better-perceived health can be ensured for everyone (Raymond et al. 2016; see Korpela et al. 2010).

A small but growing body of research suggests that alongside the well-known health risks of water bodies (e.g. drowning, WHO,2014; White et al. 2020) they also offer potential benefits for health and well-being (Finlay et al. 2015; Gascon et al. 2015; 2017; Miller et al. 2012; Völker and Kistemann, 2011; 2013; Wheeler et al. 2012; White et al. 2013; White et al. 2016, Depledge et al. 2019). Some benefits, like opportunities for physical activity, require proximal contact in, on, under, or by the water (Papathanasopoulou et al. 2016; Pasanen et al. 2019; White et al. 2014) while other benefits such as relaxation and restoration for stress (e.g. White et al. 2013; Garrett, 2019, Dempsey, 2019, Nutsford, 2016) can be gained distally/virtually (being able to see, hear or otherwise sense water).

The pathways linking blue spaces and health (White et al., 2020) are thought to be similar to those proposed for green spaces (Frumkin et al. 2017; Hartig et al. 2014; Markevych et al., 2017; Miller et al. 2012; Kabisch et al. 2017) including instoration (e.g. encouraging physical activities) and restoration (e.g. psychophysiological stress recovery). When compare blue spaces with green spaces for social benefits, in the balance of evidence blue spaces provide different recreational opportunities and mental health benefits i.e. greater levels of restoration and improved mood and feeling of freedom (Brown, 2020). Health and well-being outcomes vary according to a range of behavioural affordances (e.g. splashing, swimming, diving, walking along a river, sunbathing at a beach, etc.) depending on the type and amount of exposure.

There are frequent incidental presences of waterscapes within areas categorised as green spaces (Foley and Kistermann, 2015) and these offer unique and different experiences. Blue spaces are associated with higher landscape preferences (White et al. 2010; Völker and Kistemann, 2011) and intense perception of the environment, notably symbolic-semantic influences, contemplation, and offer a better sense of a spatial marker and atmosphere (Völker and Kistemann. 2015).

However, discussion of the potential benefits has, to date, focused on broad-brush blue space categories e.g. the coast (White et al. 2013) or inland rivers or lakes (Volker & Kistemann, 2011) rather than on the subtle differences of specific exemplars. Coastal and inland waters vary enormously, and so, we assume, might their health-promoting potentials.

The BlueHealth research project funded by the European Union Horizon 2020 research and innovation programme (https://bluehealth2020.eu/) focused on strengthening the evidence base surrounding the potential health and well-being benefits of urban blue spaces (such as docks, rivers, harbours, coastal areas, lakes, canals, and water features). One key aspect of this work was the development of a set of research and planning tools. One of these - the BlueHealth Environment Assessment Tool (BEAT for short) - was exclusively designed to map blue space factors present at a given location to help identify the terrestrial environmental features present within the site that might aid (or hinder) afford or discourage, health and well-being promoting behaviours such as exercise as well as relaxation. The BEAT was conceived as a tool that could be used by planners and designers to aid improvements, upgrades, or maintenance of blue spaces, with human health and well-being in mind.

The BEAT development was based on a ‘Person-Environment interaction’ model for blue space (Figure 1), according to the ‘person-environment fit’ theory (Mishra et al.

2020; also see Suresh et al. 2006; Korpela et al. 2008) and it establishes potential linkages between the physical attributes of blue space and the promotion of health and well-being.

It is hypothesised that the potential of blue space to support various health outcomes operate through two ontological dimensions of the blue space physical environment: environmental affordances and affect (Mishra et al. 2020). The BEAT is intended to help guide blue space physical interventions and management practices using an expert assessment of the environment. For planners and designers to have confidence in the tool, it is necessary to demonstrate that the results can be reliable and useful and that it works effectively when applied to a given blue space.

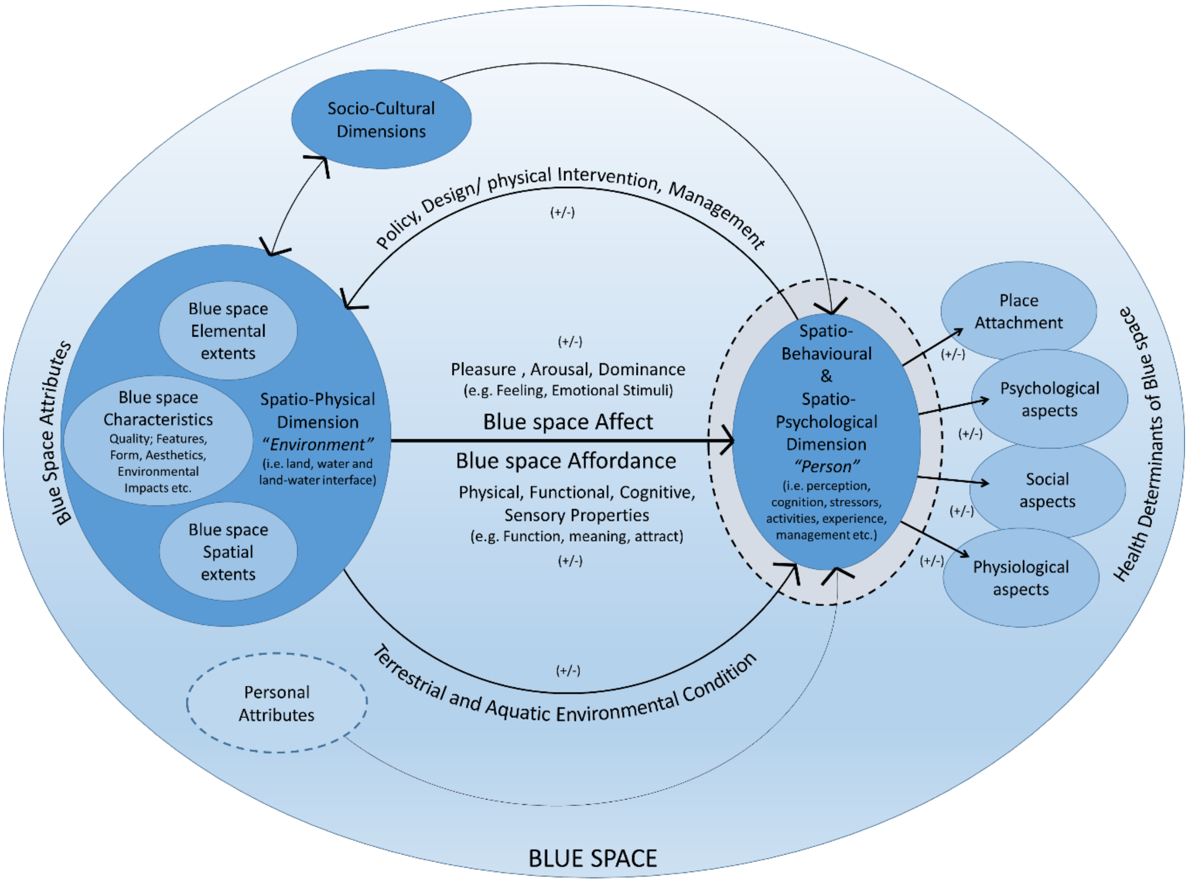

Figure 1: The Person-Environment interaction model for "Blue Space and Health" outcomes which provide the theoretical basis for the BEAT (after Mishra et al. 2020). On the left are the blue space

attributes assessed by the tool and the rest of the model shows the pathways which lead to the expected health outcomes.

The BlueHealth Environmental Assessment Tool (BEAT).

The BEAT tool (Mishra et al. 2020) is freely available at https://www.beat.bluehealth.tools/. The tool website includes several components. Firstly, it contains the BEAT tool. This comprises a four-step workflow: three of these steps pertain to the terrestrial part of a blue space while the fourth step (Ott et al. 2017) facilitates an assessment of aquatic ecology. There are two versions of the tool: one for use by professionals and one for use by communities in the context of citizen science. Secondly, the website features guidance documents and other materials such as a compendium of images of a wide range of blue space types. The three steps of the professional version of the terrestrial BEAT are (1) site survey and data collection, which is a desk study to gather factual information about the area to set the specific site in its wider context-macro-level assessment; (2) qualitative information formed by the first impression of visiting the site; and (3) the site-level collection of data for a comprehensive assessment of a range of site attributes - micro-level assessments, using various scoring systems, primarily a 5-point Likert-type rating scale applied to all attributes (when they are present). The assessments can be carried out using the online tool directly or a paper version of the tool forms, after which scores can be uploaded to the database, and the results can be downloaded by assessors using their access codes.

The recommended method for applying the tool is for two assessors to carry out the survey independently and then to pool the results to test the inter-rater reliability of their assessments. It is likely that some attributes (e.g. aesthetic, social, items with temporal variability) are more difficult to rate, are rated inconsistently over time, or their qualitative nature might result in greater differences (Brownson et al. 2004; Saelens et al. 2006). Therefore, deeper training, setting reliable indicators, and repeated assessments over time could reduce subjective influences on the rating. In addition, moderation of any divergent scores through discussion by the assessors can be useful for a specific purpose in a given circumstance. These discussions may be key to revealing interesting perspectives on the place and are an important aspect of the application of the tool.

This paper focuses on Step 3, the on-site survey of the BEAT (professional version). Ideally, a tool of this kind also ought to be sensitive to real differences between sites. The blues space attributes that are assessed in the BEAT step-3 were expected to provide an accurate assessment of physical, social, and aesthetic features of terrestrial blue space which are potentially important for promoting health and well-being. The BEAT examines the combined influences of blue space features and provides an informed, balanced evidence base by gathering data on the quality and characteristics of physical, social, and aesthetic aspects of the same place, rather than using presence or absence criteria for the assessment of potential affordances for physical activities and health.

Many studies have demonstrated and tested the inter-rater reliability of similar tools which assess the physical environment of, for example, neighbourhood green spaces, parks, and playgrounds e.g. Neighbourhood Green Space Quality Assessment Tool (NGST) (Gidlow et al. 2012); environmental Assessment of Public Recreation Spaces Tool (EAPRS) (Saelens et al. 2006); streets, walking and cycling paths e.g. Walking and Bicycling Suitability Assessment (WABSA) (Emery et al. 2003), Audit Tool Checklist Version (ATCV) and analytical Version (ATAV) (Brownson et al. 2004); characteristics of trails and paths e.g. Path Environment

Audit Tool (PEAT) (Troped et al. 2006) and their potential to promote physical activity and healthy behaviours. However, to our knowledge, only a few studies have focused on water as an element in a park setting. For example, Saelens et al. (2006) reported an analytical assessment of cleanliness, water quality, accessibility to the water and specific qualities, the assessment of proximity and place dimensions using a 3-point rating scale. For both initial and second observation, the results suggested good to excellent reliability for the criteria i.e. presence/absence and number of items, and specific qualities observed at the water areas. Other studies reporting reliability mostly record the presence/number of items (e.g. pool), presence and visibility of types of recreational facilities (e.g. outdoor pool, beach, marina, etc.), and natural features (e.g. large or small water bodies such as the ocean, lake river, pond, stream, etc.) (Tropped et al. 2006; Gidlow et al. 2012). Even though Cavanar et al. (2004) included assessment items e.g. availability of specific aquatic facilities and safety aspects, they did not report reliability.

Environmental audit tools are intended for research purposes and/or to support local decision making (Brownson et al. 2009). However, following best practice, the BEAT was developed with the potential to be used not only by professionals and researchers but also by communities (in the spirit of citizen science) and proposes an in-depth and detailed assessment of any blue space. This paper focuses on the reliability and feasibility of the professional version of the BEAT following the best practice approaches developed earlier for environmental audit tools such as the Environmental Assessment of Public Recreation Spaces Tool (EAPRS) (Saelens et al., 2006) or the Path Environment Audit Tool (PEAT) (Troped et al. 2006). These tools provided in-depth guidance notes regarding the structure, utility and operational aspects of the tools to increase tools' effectiveness and reliability (Brownson et al. 2009).

The current study aims to test the reliability of the BEAT tool in assessing different types of blue space, in terms of inter-rater reliability (consistency of evaluation between different assessors) and its effectiveness (how well it identifies key health-promoting affordances in a given blue space). We, therefore, wanted to know: (a) how the survey tool performs when used by different assessors at different sites, and (b) how effective the tool is at discerning real differences between site quality and characteristics that may inform differences in potential for health promotion.

Materials and Methods

We adopted two methods (Figure 2) to answer the research objectives: (1) the inter-rater reliability testing in two distinct stages (pilot-testing at Stage-1 and reliability testing at Stage-2) on two different sets of test sites for objective (a); and (2) the effectiveness and variability testing for objective (b) using the Stage-2 set of sites. In the reliability testing process, we introduced deeper training at Stage-2 to test the effect of this on improving the reliability of scores for the attributes following the pilot testing results from Stage-1. Figure 2 shows the general approach we adopted.

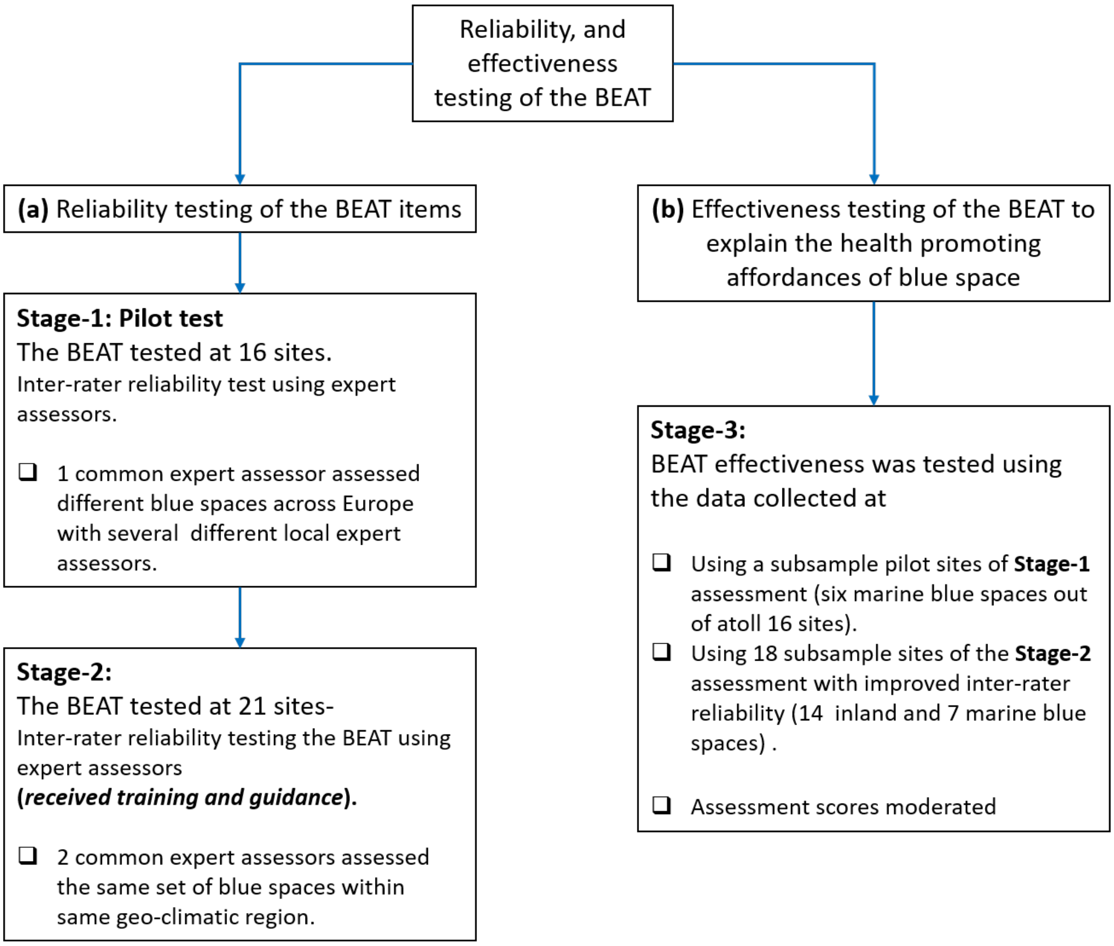

Figure 2: The general strategy adopted in the research.

Assessment components

The BEAT Step 3 is broadly structured according to physical, social, aesthetic, and environmental domains, each being sub-divided into several aspects and each aspect containing a variable number of attributes or items (Table 1; Supplementary Material, Appendix 1, Table 1; also see Mishra et al. 2020). The blue space attributes from 16 aspects were assessed and rated during the on-site survey are listed in Table 1. All 16 aspects were assessed for inter-rater reliability at Stages 1 and 2, and 10 aspects i.e. access and circulation- condition, access and circulation-visual appearance, access and circulation- functionality, disabled access, site management, information and education, safety and security, visual condition, visual quality and non-visual aesthetics were tested for the BEAT's effectiveness in Stage-3. The total number of possible attributes that could be assessed was 125, structured within 16 aspects (see Table 1). Not all of these were scored at every site because not all features are present at all sites. Thus, in some cases, no comparison could be carried out between Stages 1 and 2.

Table 1: The number of aspects and attributes assessed in Stage-1 and Stage-2 testing (Refer Supplementary Material Appendix 1, Table 1 for the complete list of attributes for the aspects listed.).

Domain (N-3)	Aspects (N=16)	Number of attributes assessed (N-125)
Physical	1. Access and Circulation (condition) *i.e. access roads within the site; car parking onsite; path constriction and use of material etc.*	6
	2. Access and Circulation (visual appearance)	6
	3. Access and Circulation (functionality)	6
	4. Disabled Access *i.e. physical disabilities access conditions*	4
	5. Terrestrial Recreation Structure (condition) *i.e. toilets, changing rooms, food, and ice-cream store, etc.*	12
	6. Terrestrial Recreation Structure (visual appearance)	12
	7. Terrestrial Recreation Structure (functionality)	12
	8. Water Recreation Structure (condition) *i.e. boat slipway; jetty, pier, etc.*	10
	9. Water Recreation Structure (visual appearance)	10
	10. Water Recreation Structure (functionality)	10
	11. Site Management *i.e. maintenance of hard surfaces; management of vegetation, furniture, etc.*	6
Social	12. Information and Education *i.e. presence, and usefulness of information; functionality of waymarking or directional signs, etc*.	6
	13. Safety and Security *i.e. presence of water safety equipment and lifeguards; presence and functionality of lighting etc.*	7
Aesthetic	14. Visual Condition i.e. Visual quality of buildings and other structures; Sense of openness and scale of water views etc.	6
	15. Visual Quality i.e. Quality of views within the site; Attractiveness of vegetation on the site etc.	6
	16. Non-visual Aesthetics Qualities *i.e. pleasant smells within the site; absence of unpleasant smell; pleasant sounds within the site, etc.*	6

All attributes were graded using a rating scale (Likert-type scale: 0-not present or not relevant, 1 for "very low" to 5 for "very high", according to the context of the attributes and the points mentioned in the guidance). In the Stage-1 pilot testing for reliability, assessments were carried out with limited training and without necessarily closely following the guidance notes. Later in the Stage-2, assessors were given deeper training and asked to follow the guidance notes in detail during the assessments. For example, when rating the "Quality of views out from the site across the water" to explain the "visual condition of site surroundings", one must answer the following questions (1) "When standing on the water's edge are the views looking towards attractive features, natural elements present, historic structures or the open horizon?"; (2) "Are there features such as derelict sites, unattractive buildings or buildings in poor condition visible?" The option to take notes during the assessment is designed to capture qualitative information and to help surveyors to moderate their final scores and for qualitative analysis. This is important because the scoring system only forms part of the whole BEAT system.

The tool includes a section for recording the users of the site at the particular time the survey was conducted. As both assessors may not have carried out their survey synchronously, such information is time-specific and so no consistency between their results for these sections would be expected. Thus, these sections were removed from the data used in the current study.

Selection of test sites

We selected two sample sets of urban blue spaces representative of different types. To test the reliability of the tool, for the pilot testing Stage-1, the sample comprised 16 urban blue spaces located in five European countries: Estonia (n=8), Finland (n=4), The Netherlands (n=2), the UK (n=1), and Spain (n=1). For Stage-2 the sample comprised 21 urban blue spaces located within the two main cities of Estonia: Tartu (n=10) and Tallinn (n=11) (Figure 3 and Table 2 a & b).

The effectiveness of the BEAT was tested using data from the assessment stage-1 (i.e. pilot sites with low training and guidance) and Stage- 2 (i.e. test sites with deeper training and guidance). To test the ability of the BEAT to explore the health-promoting affordances at Stage 3, we selected a sub-sample of six coastal pilot sites which were grouped into three site-types i.e. coastal waterfront; coastal waterfront-beach; and bayfront (Table 2a) from the Stage-1 sample sites (n=16) located in Estonia, the UK, and Finland. From Stage-1 sites, only coastal sites were selected as sub-sample sites to see whether the BEAT is able to pick up intra-site (i.e. between sites within each site-types) and inter-site-type differences (i.e. between site-types) when applied within a similar environmental context (i.e. all marine environments) with minimum guidance and training support. Further, to test the effectiveness of the BEAT to capture the intra-site and inter-site-type differences when applied to different blue environmental settings, we selected a sub-sample of sites (n=18) from a wide range of coastal and inland sites (Table 2b) from the Stage-2 sample sites (n=21; Estonian sites) (three were omitted because they could not be compared very easily) with deeper training and guidance.

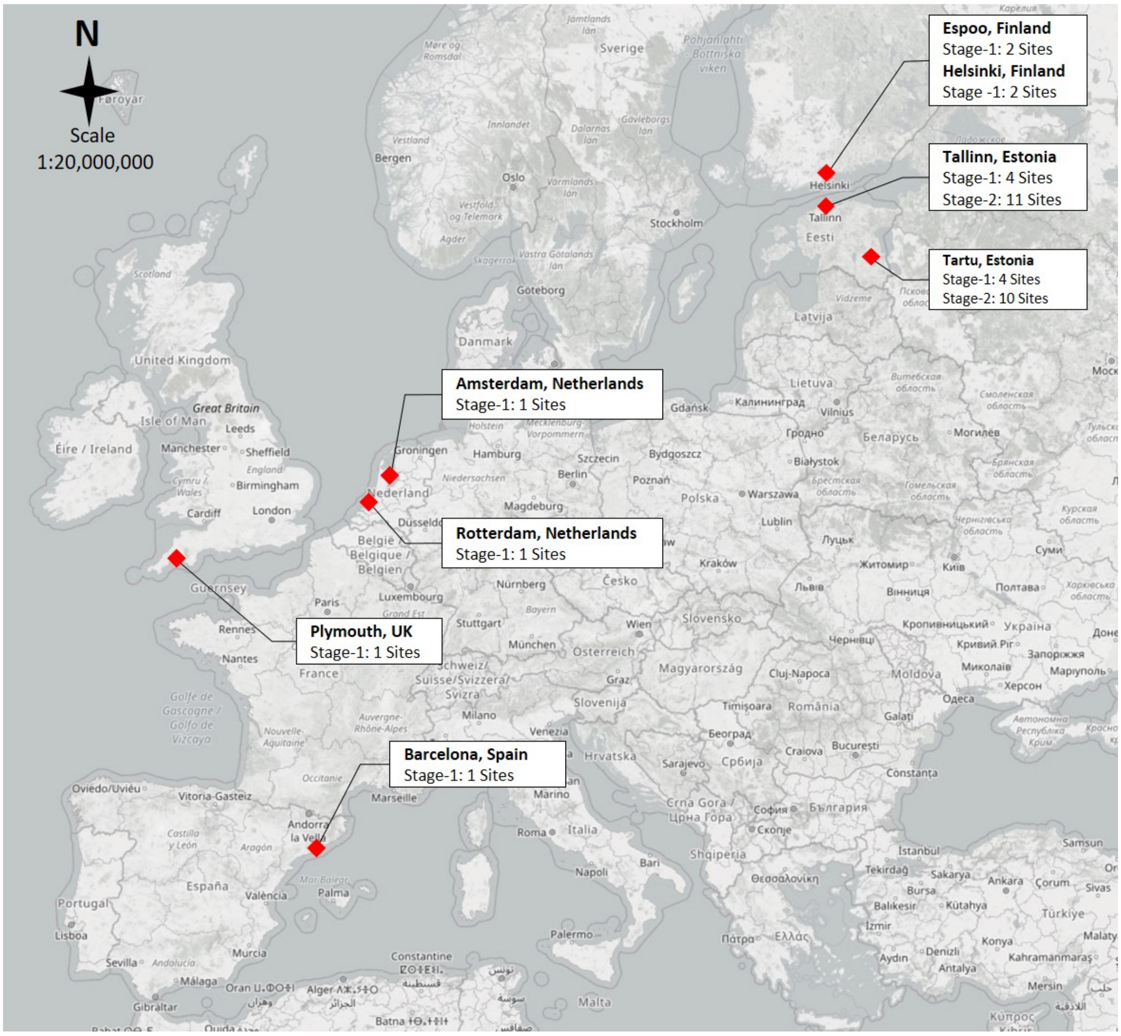

Figure 3: Location of sites assessed for Stage-1 and Stage 2 testing, (Source: OpenStreetMap - © OpenStreetMap (and) contributors, CC-BY-SA).

We selected the Stage-1 pilot and Stage-2 test sites according to five criteria: (1) they featured different socio-cultural characteristics; (2) they were moderate in territorial scale and located within or next to an urban neighbourhood; (3) they were publicly accessible and used for formal and informal recreation; (4) they represented different geographical settings and ecosystems; (5) they represented different behaviour settings (see Table 2 a and b).

We classified sites based on built character and degree of naturalness into three site-types for Stage-1: (1) coastal waterfront (sites 1 and 2), (2) coastal beach (sites 3 and 4), (3) coastal bayfront sites (sites 5 and 6) (see Table 2a) ; and nine site-types for the Stage-2: (1) River-beach natural (sites 17 and 18), (2) River-embankment walk- natural, built (sites 19 and 20), (3) Canal- beach, natural bank (sites 21 and 22), (4) Fountain- built (sites 23 and 24), (5) Small lake- park, natural (sites 25 and 26), (6) Dockland- semi-natural, built (sites 27 and 28), (7) Coastal-beach (sites 29 and 30), (8) Bay-cliff, beach (sites 31 and 32), (9) Pond-park (sites 33 and 34) (see Table 2b).

Table 2: The location and type of the samples used in Stage-1 and Stage-2 testing

(a) Blue spaces tested at Stage-1				
S.N.	Site-types	Site name	Blue space type	Environment type
Coastal				
1	Coastal waterfront	Teats Hill, Plymouth, UK	Sea- Stony beach, marina	Urban green space character by decline and deprivation.
2		Pelguranna Beach, Tallinn, Estonia	Sea- Stony beach, harbour	Undeveloped space with Industrial character
3	Coastal waterfront-beach	Matinkyla Beach, Espoo, Finland	Sea- Stony and sandy beach	Natural setting with urban facilities.
4		Kopli Beach, Tallinn, Estonia	Sea-Sandy beach	The natural character next to high-density housing with urban facilities.
5	Bayfront	Elaintarhanlahti Bay, Helsinki, Finland	Bayfront- Marina and promenade	Linear greenspace facing the bay with urban facilities and within a high-density urban district.
6		Otsolahti Bay park and marina, Espoo, Finland	Bayfront- Marina, and promenade	Large green open space and marina within a natural setting.
Inland				
7	Riverfront	Besos river along Montacada i Reixach, Barcelona, Spain	Large river with artificial banks embankment	Large fluvial park with embankment walk within a high-density urban district.
8		Emajõgi River Walk, Tartu, Estonia	Medium-sized river with natural banks,	Green embankment walk with a natural setting within low-density development
9	Large pond	Väike-Õismäe pond, Tallinn, Estonia	Pond- Riprap banks	Green open space within high-density housing.
10		Tapiola Kulturikeskus lake, Espoo, Finland	Pond- Artificial banks	Open space within a high-density commercial district.
11	Lake	Päe park, Tallinn, Estonia	Small Lake- Artificial banks with a cliff	Large green open space around an old stone quarry within high-density housing.

12		Meer en Vaart Boulevard, Amsterdam, Netherlands	Large lake-Dock and constructed promenade	Artificial and natural lakefront within high-density community district with urban facilities and services.
13	Canalised waters	Spoorsingel, Rotterdam, Netherlands	Urban canal-Artificial banks	Canalised urban waterbody with a natural bank and dense vegetation and within a high-density urban district.
14		Anne Kanal, Tartu, Estonia	Artificial lake and River with a natural ba	Urban natural linear green space between a river and artificial lake next to a high-density urban housing district.
15	Ornamental water features	Raekoja Plats, Tartu, Estonia	Ornamental water feature or fountain	A city centre urban plaza with a fountain.
16		Tartu University library plaza, Tartu, Estonia	Ornamental water feature or fountain	Urban open space with ornamental water features within a highly built-up residential and institutional setting.

(b) Blue spaces tested at Stage-2

S.N.	Site-types	Site name	Blue space type	Environment type
17	River beach-natural	Emajõe vabaujula, Tartu, Estonia	Medium-sized river with natural banks	Sandy river beach on the outskirts of town, some bushes, high usage in the summer season.
18		Emajõe linnaujula, Tartu, Estonia	Stream with natural banks	Sandy riverside beach. People use it for walking and relaxing.
19	River embankment - natural, built	Emajõe walk, Tartu, Estonia	Medium-sized river with natural banks	Riverside alley of poplar trees, used for walking, running and relaxing, fishing.
20		Emajõe City centre walk, Tartu, Estonia	Medium-sized river with artificial banks	Riverside artificial promenade with some seasonal cafeterias. Mainly used by pedestrians.
21	Canal- beach, natural bank	Anne Kanal Beach, Tartu, Estonia	Urban artificial lake next to a residential area	Sandy beach of an artificial water body, used for swimming and sunbathing.
22		Anne Kanal Island, Tartu, Estonia	Urban artificial lake next to a residential area	A pedestrian path is surrounded by grassland, trees, and bushes used by pedestrians.
23	Fountain-built	Raekoja Plats, Tartu, Estonia	Ornamental water feature or fountain	Iconic fountain in the central square of the old town. Used a lot by tourists. Restaurants, shops, offices.
24		Tartu Library Plaza, Tartu, Estonia	Ornamental water feature or fountain	Town plaza in the font of University Library, used for resting passing by.
25	Small lake-park, natural	Päe park, Tallinn, Estonia	Artificial lake	Artificial lake in a park, used for recreation.
26		Harkujarv, Tallinn, Estonia	Natural lake	Lake with the sandy beach near residential areas. Popular in summer.

27	Dockland- semi-natural, built	Kalarand, Tallinn, Estonia	Old dockland and sea	The area is used for informal activities with a small beach. It used to be a fishing harbour and now is an urban wasteland
28		Sadama Turg Tallinas, Tallinn, Estonia	Dockland	Harbour promenade. Used for harbour and pedestrian use.
29	Coastal- beach	Pikakari Rand, Tallinn, Estonia	Sea	Sandy beach with an old dock edge now serving as a seaside promenade.
30		Pirita Rand, Tallinn, Estonia	Sea	Sandy beach, used for swimming, sunbathing, and windsurfing.
31	Bay- cliff, beach	Kopliranna, Tallinn, Estonia	Bay	Abandoned industry site, currently not in use, tall grass, natural paths.
32		Pelguranna beach, Tallinn, Estonia	Bay	Large sandy beach with the recreational park, used for swimming, sunbathing, etc.
33	Pond- park	Memorial statue of Tartu Mayor Karl Luik, Tartu, Estonia	Pond	Town park with water features and monuments. Used for relaxing.
34		Õismäe Tigi Park, Tallinn, Estonia	Pond	Artificial symmetrically laid out park in the middle of a residential area.
35	Others	Ülejõe park, Tartu, Estonia	Medium-sized river with artificial banks	Park in the city centre.
36		Russalka Rand and promenade, Tallinn, Estonia	Sea	Seaside promenade for pedestrian use, cycle path.
37		Snelli pond, Tallinn, Estonia	Artificial stream	Park in the edge of the old town used for leisure and strolling. Many tourists.

Site survey and data collection process

Following best practice (Brownson et al. 2009; Zhang et al. 2017), sites were surveyed by two experts. During both stages, the experts carried out assessments separately and independently. Also during both stages, on-site survey practical aspects were discussed such as procurement of area maps, desktop study, etc.

For Stage-1 we recruited local experts from within the field of landscape architecture, urban planning, and landscape management who were locally available, were familiar with local context and history e.g. town or region, and had been involved in planning, design, or management of similar local landscape projects or setting. Some of these were identified by BlueHealth research partners, others were colleagues of the research group in Estonia who were not involved in the BEAT development. For the Stage-1 pilot testing, each site was assessed by different local assessors and one common reference assessor. The primary author of the study (one of the developers of the BEAT tool) acted as the primary and reference assessor for all 16 sites. Each of the secondary assessors then rated one or more of the sites independently. From a total of 16 sites, the secondary assessors visited 14 sites on the same day and two sites within the same month. The Stage-1 assessments took place during July 2017 for Plymouth, UK; October 2017 for Tartu and Tallinn, Estonia; November 2017 for Barcelona, Spain; September 2018 for Espoo and Helsinki, Finland; and October 2018 for Amsterdam and Rotterdam, Netherlands. In the absence of a "gold standard" blue space reference case for comparison (Zhang et al., 2017), we used the assessment scores of the primary assessor to create a proxy standard for all sites. All sites were surveyed during daylight hours and in favourable weather and the assessors focused on the constant characteristics of the physical environment. Thus, we did not consider possible influences of light, weather conditions or cultural landscape on the scores.

For the Stage-2 testing, two new experts assessed all the sites. Experts rated every site at the same time (though separately) between October and November 2019. In this way, the consistency of scoring could be tested and some of the lessons learned in Stage-1 could be applied to improve the inter-rater reliability. The assessors for different stages, received different levels of guidance and training.

Following best practice (Bedimo-Rung et al., 2006; Saelens et al., 2006), before carrying out the Stage-1 assessment the local assessors received some training on the content of the BEAT and the operation of the BEAT online interface for a tablet or a smartphone. The Stage-1 pilot testing was carried out as part of the BEAT development. The tool was initially developed by a group of landscape architects and later applied by other experts who were architects, landscape architects, landscape managers or planners. Thus, we were interested in their feedback about the quality of the guidelines, functional aspects of applying the tool, the clarity and relevance of the questions in order to improve the guidance notes and to decide on the degree of in-depth training to give the assessors for Stage-2. For the Stage-1, both assessors performed the on-site task using paper forms and later submitted the scores into the online BEAT forms which fed the results into the BEAT online database. The Stage-2 assessors used the BEAT online system for the survey.

During Stage-2 operational definitions were established and the draft guidelines and onsite survey instructions were provided which were common for all the sites. For Stage-2, the assessors attended an in-depth training session that included an introduction of the BEAT and its contents, detailed explanations about the blue space aspects and attributes and hands-on outdoor training on how to use the BEAT online tool using a tablet or smartphone.

All assessors received guidance on points to consider while answering each question to avoid problems regarding the identification of features or interpretation of a specific condition on-site and how to interpret more subjective attributes. Prior to the site assessments, assessors were advised to carry out a complete site inspection, a visual survey of the surrounding area e.g. neighbourhood characteristics or housing quality for each site and to agree the common routes and points of observation. They were recommended to divide a larger site into sub-zones based on different biotopes or ecological variability, to identify all access points and how to manage the task within a specified time period.

The BEAT contains sections for assessors to record in note form their subjective opinion about any attribute they assess. However, the use of the subjective data to supplement the scores lies beyond the scope of this paper. They were also asked to make notes about the practical application of the tool, any problems they encountered or technical issues. Each expert could use online supplementary guidance notes during the on-site assessments and they submitted their scores online using a tablet computer.

The BEAT provides both the option of using either the online BEAT interface, or printable paper forms (for use where wifi connections may be poor or the aim is for both assessors to discuss their scores and comments before data is entered). For example, in Step-1 the assessors used the paper forms and found it easier to make more extensive notes about many of the attributes during the assessment. This was not only additional explanatory information but also comments on the functionality. Notes also help to justify the ratings in order to produce an agreed score for an attribute of a greater degree of subjectivity by using a moderation process, and for filling in missing data and scores incorrectly entered.

Data editing and analysis

Using an earlier established protocol (Brownson et al., 2004), we carried out cleaning and organising of the data which had been submitted to the online system. The data were downloaded as a spreadsheet using the .xlsx format. The 5-point Likert-type scale scores collected during both stages of assessment were structured according to 16 categories (i.e. 16 files containing each aspect or sub-aspect: eg. *information and education, safety and security, etc.)* for all the sites. The files were subsequently reviewed for missing codes or repetition; any missing scores per attribute per assessor were identified and the relevant experts were contacted to supply revisions by reference to their original paper forms used for Stage-1 and personal notes taken during Stage-2. Out of the total number of BEAT items, at Stage-1 55.15%, and Stage-2 48.91% were present on the sites and assessed. Data cleaning (Willes, 2017) here means reviewing and fixing common errors i.e. entering a wrong site code against attribute scores and any post-moderation changes to the ratings. Out of the total items assessed, in Stage-1, 8% and in Stage-2, 7.81% were missing or incorrectly entered. The time required to complete the survey for each site (i.e. terrestrial environmental assessment only) by the experts generally took between three to four hours depending on its size of the site, including time spent walking around to get a feel for it, checking each aspect in detail, filling out the paper or online forms and taking photographs (Mishra et al., 2020). The scores were then subjected to quantitative data analysis. All analyses were conducted using IMB SPSS Statistics version 26 (IMB Corp. 2019).

Calculating the inter-rater reliability and agreement.

We tested the inter-rater reliability of scores for Stages 1 and 2 using interclass correlation (ICC: an absolute agreement type, two-way random effect model). ICC is commonly used to assess the variation between two or more raters who measure the same group of subjects (Koo and Li, 2016). We tested the reliability of the BEAT attributes based on the scores given on a Likert-type scale (1-5) annotated a set of explanatory notes in the BEAT guidance for different conditions of the attributes. To achieve a robust and reliable assessment tool the test of the absolute agreement is the best measure when a narrow scale range is used to assess the quality of a feature in an environment (i.e. 1 to 3 or 1 to 5, given that the consecutive scores on the scale may suggest a very different environmental quality or characteristics of the same attribute). We tested the internal consistency of item scores using Cronbach's alpha (acceptable at $\alpha \geq 0.7$) (Taber, 2018). For calculating the overall degree of agreement between assessors we estimated Fleiss Multi-rater Kappa coefficients (k) for the assessment scores for categorical assessment items - blue space attributes for aspects e.g. access and circulation, site management, safety and security, visual condition, etc. (see Supplementary Materials Appendix 1, Table 1). As a guide to interpreting the results of the ICC for the reliability and k-values for the agreement between the two observers we used ratings in the following categories: < 0.20 (poor); 0.21- 0.40 (fair); 0.41- 0.60 (moderate); 0.61- 0.80 (substantial) and 0.81- 1.00 (almost perfect) (Landis and Koch, 1977).

Testing the effectiveness of the BEAT to explore health-promoting affordances

For testing the effectiveness of the BEAT in terms of its ability to identify features that could promote (or hinder) health-promoting behaviours, we used the moderated scores of the two observers for the selected sub-sample of sites for both the stages (see Table 2 a & b). Descriptive statistics (frequency distributions) and visualisation of multi-variate data (i.e. domains-aspects-attributes) were used to summarise differences in the condition and quality of different place aspects which represent the baseline conditions of the sites. Spidergrams were used to plot the values of all the attributes within each aspect, they were useful to illustrate and compare intra-site variations and thus the sensitivity of the BEAT to discern site-specific differences and unique characteristics. A Shapiro-Wilk normality test ($p<0.05$) confirmed that the scores were non-parametric (Mishra et al., 2019). A Kruskal-Wallis H (one-way

ANOVA; $p < 0.05$) test (McKight and Najab, 2010) was carried out to report the significance of intra-site (between sites within a site type) and inter-site (between all the sites) variance of attribute scores for both the sub-sample sets grouped by categories (site-types, aspect-types and domain-types), which were observed in the spidergrams. A spidergram is a useful graphic tool to interpret values of a specific aspect that varies over different attributes or parameters. To interpret results, on a chart, each axis represented a category that received a score on a Likert-type scale (i.e. 0- not present or not relevant, 1 for "very low" to 5 for "very high"). A concentric ring that connected the points located on each axis created a shape that explained the categories that stood out. These are visually clear and easy to compare when different scores are seen together within a single spidergram or where several spidergrams are placed side-by-side.

Results

Inter-rater reliability for BEAT items at Stages 1 and 2

ICCs were produced for 82.4% and 76.0% of the total possible attributes assessed at Stages 1 and 2 respectively (owing to the type of sites at Stage-2, there were more possible site elements missing.). Out of a total number of attributes analysed for ICC values, 65.04% for Stage-1, and 82.10% for Stage-2 were found consistent at an acceptable level (Cronbach's alpha values (α) ≥ 0.7). In the Stage-2 assessment, 53.6% of the tested items had reliability values in the "almost perfect" range for ICC- absolute agreement values compared to 44% in the Stage-1 assessment. The complete record of inter-rater reliability for Stages 1 and 2 according to the different ranking classes is presented in detail (with each ICC and k-value) in the (see Supplementary Materials Appendix 2, Table 1 and 2) and key findings summarised below. Missing ICC values for attributes indicated too few cases were present or no cases were calculated. Any negative ICC values suggested poor or low agreement between raters due to the small sample size (n) (Giraudeau,1996; Liljequist et al., 2019).

The two-stage reliability testing revealed an improvement in scores from Stage-1 to Stage-2 for several attributes that required a more subjective interpretation of place quality and characteristics. For example, in the social domain, notable improvements in ICC values were found for “functionality of way-marking or directional signs” and “accessibility information for people with disabilities.” Similarly, a notable improvement in inter-rater reliability in Stage-2 was observed for items assessed for aspects such as “visual condition”, “visual quality” and “non-visual aesthetics” and “disabled access.” The ICC agreement values calculated for the items of “site management” conditions for Stages 1 and 2 varied less. Overall, at Stage-2 the ICC values for the assessment scores improved for the same items with subjective influences calculated in Stage-1. Improvements from “substantial (0.61-0.80)” to “almost perfect (0.81-1.00)” ICC agreement values were noted for 15 items in Stage-2 out of the same 22 items in Stage-1, such as “sense of openness and scale of water views”, “screening of eyesores”, “sense of wildness” and conditions related to “vandalism and lack of safety.” The 12 Items with “moderate” ICC values in Stage-1 also improved to “substantial (0.61-0.80)” and “almost perfect (0.81-1.00)” in Stage-2, such as attributes related to “management of vegetation”, conditions related to physical disability access, smell, and the visual quality of built structures and quality of views within the site. Similarly, all items with “fair (0.21-0.40)” ICC agreement were improved at Stage-2 to “substantial (0.61-0.80)” and “almost perfect (0.81-1.00)” values for attributes with high health-related values such as “feeling of tranquillity”, "attractiveness of vegetation and sense of atmosphere." Items with "poor (< 0.20)" ICC agreement values such as "olfactory condition (pleasant smell)" and "sense of general security against crime and anti-social behaviour" were improved to "almost perfect" ICC agreement values at Stage-2.

A majority of the items for visual appearance and functionality of terrestrial and water recreational structures already showed moderate to excellent ICC agreement values at Stage1 with no change at Stage-2, except for the items for elements such as toilets, fountains, art installations, safety equipment store, food and ice-cream stalls and bridges. For aspects within the physical domain, notable improvements in ICC agreement were chiefly observed between Stag-1 and Stage-2 for items such as "condition of footpath network" ICC = 0.16 (poor) – 0.96(almost perfect); "functionality of footpath network" ICC = 0.20 (fair) – 0.91 (almost perfect); "functionality of the use of path materials" ICC = 0.23 (fair) – 0.87(almost perfect); and "maintenance of street furniture" ICC = 0.35 (fair) – 0.86(almost perfect). For the social domain, large improvements were found for the aspects "functionality of way-marking signage" ICC = 0.29 (fair) – 0.92(almost perfect); "site information in different languages" ICC = 0.44 (moderate) – 0.94(almost perfect) within "information and education" and "security against crime or anti-social behaviour" ICC = 0.04(poor) – 0.86(almost perfect) within the aspect of "safety and security". For the aesthetic domain, reliability improved for attributes within the "visual quality" aspects such as "attractiveness of vegetation" ICC = 0.25 (fair) – 0.85 (almost perfect); and for the "non-visual aesthetic" aspects such as "sense of atmosphere" ICC = 0.20 (fair)– 0.73 (substantial) or "feeling of tranquillity" ICC = 0.29 (fair) – 0.87(almost perfect). Generally, the range of confidence interval (CI) of ICC values per attribute has improved from Stage-1 to Stage-2.

For the Stage-1 and 2 assessments, out of all the items assessed, the overall agreements between two assessors calculated using Leiss Multi-rater Kappa coefficients (k-value) were found to be 5.6% and 9.6% for “almost perfect (0.81- 1.00)” to “substantial (0.61- 0.80)”, 42.5% and 47.2% for “moderate (0.41- 0.60)” to “fair (0.21- 0.40)”, and 12.8% and 20.8% for “poor (< 0.20)” range, respectively. Fewer items in Stage-2 (22.4%) when compared to Stage-1 (27.2%) showed no agreement. For Stage-2, the overall agreement was improved for different ranking classes. Overall, the k-values for the Stage-2 assessment showed improvement in the items for aspects of “access and circulation”, “safety and security”, “information and education”, “visual condition”, “visual quality”, “access for disabled people”, and “site management.”

The effectiveness of the BEAT to explore health-promoting affordances (Stage 3)

Table 3: Kruskal Walis H test (one-way ANOVA) of a non-parametric test of significance of variance of the BEAT site assessment scores for Stage-1 and 2.

(a) Kruskal Wallis H Test For Sites Assessed in Stage-1

	Grouping Category: Site types			Grouping Category: Aspect-types			Grouping Category: Physical domain			Grouping Category: Social domain			Grouping Category: Aesthetic domain		
	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)
Intra-site variance[a]															
1. (site 1, site 2)	15.495	1	0.001	31.775	9	0.001	3.053	4	0.549	9.298	1	0.002	3.879	2	0.144
2. (site 3, site 4)	8.253	1	0.004	19.026	9	0.025	10.885	4	0.028	0.198	1	0.656	0.661	2	0.719
3. (site 5, site 6)	0.258	1	0.611	74.177	9	0.001	22.605	4	0.001	11.463	1	0.001	8.729	2	0.013
Inter-site variance[b]															
All six coastal sites (1-6)	60.207	5	0.001	74.477	9	0.001	25.83	4	0.001	11.164	1	0.001	9.279	2	0.010
All 16 sites	74.434	15	0.001	111.429	9	0.001	19.502	4	0.001	28.235	1	0.001	5.651	2	0.059

Note: *p* significant at < 0.05 level

(b) Kruskal Wallis H Test For Sites Assessed in Stage-2

	Grouping Category: Site types			Grouping Category: Aspect-types			Grouping Category: Physical domain			Grouping Category: Social domain			Grouping Category: Aesthetic domain		
	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)	Kruskal-Wallis H	df	Asymp. Sig. (*p*)
Intra-site variance[c]															
1. (site 17, site 18)	2.317	1	0.128	30.639	9	0.001	25.464	4	0.001	5.690	1	0.017	1.617	2	0.445
2. (site 19, site 20)	12.142	1	0.001	34.045	9	0.001	21.057	4	0.001	0.486	1	0.486	2.476	2	0.290
3. (site 21, site 22)	0.008	1	0.927	37.787	9	0.001	19.586	4	0.001	5.342	1	0.021	1.780	2	0.411
4. (site 23, site 24)	2.243	1	0.134	22.629	9	0.007	7.999	4	0.092	1.817	1	0.178	1.819	2	0.403
5. (site 25, site 26)	16.767	1	0.001	10.596	9	0.304	5.544	4	0.236	0.218	1	0.641	7.002	2	0.030
6. (site 27, site 28)	10.632	1	0.001	6.406	9	0.699	1.888	4	0.756	0.840	1	0.359	2.155	2	0.340
7. (site 29, site 30)	0.590	1	0.443	19.827	9	0.019	17.634	4	0.001	0.607	1	0.436	2.524	2	0.283
8. (site 31, site 32)	4.972	1	0.026	7.243	9	0.612	2.755	4	0.600	0.156	1	0.693	1.664	2	0.435
3. (site 33, site 34)	3.061	1	0.080	28.18	9	0.001	5.469	4	0.243	3.514	1	0.061	5.865	2	0.053
Inter-site variance[d]															
All 21 sites (17 - 37)	104.412	20	0.001	106.319	9	0.001	51.788	4	0.000	11.657	1	0.001	18.074	2	0.001

Note: *p* significant at < 0.05 level

Note:

[a] Intra-site variance analysis of three site-types (Stage-1); [b] Inter-site variance analysis of all six coastal and all 16 stage-1 sites- see Table 2a; [c] Intra-site variance analysis of nine site-types (Stage-2); [d] Inter-site variance analysis of all 21 Stage-2 sites- see Table 2b,

Stage-1: site-types – (1) coastal waterfronts (sites 1 and 2); (2) coastal waterfront-beach (sites 3 and 4);(3) bayfront (sites 5 and 6); **Stage-2: site-types**- (1) river-beach, natural (sites 17 and 18); (2) river-embankment walk- natural, built (sites 19 and 20); (3) canal- beach, natural bank (sites 21 and 22); (4) fountain- built (sites 23 and 24); (5) small lake- park, natural (sites 25 and 26); (6) dockland- semi-natural, built (sites 27 and 28); (7) coastal- beach (sites 29 and 30); (8): bay-cliff, beach (sites 31 and 32); (9) Pond-park (sites 33 and 34).

Stage 1 and 2: aspect-types- Physical domain: (1) access and circulation (condition); (2) access and circulation (visual appearance), (3) access and circulation (functionality), (4) disabled access (5) site management, Social domain: (6) information and education, (7) safety and security, Aesthetics domain: (8) visual condition, (9) visual quality, (10) non-visual aesthetics.

For the sub-sample of Stage-1 sites used for this testing, Inter-site variability was statistically significant ($p < 0.05$) for all 16 sites except for the aesthetic domain (which was not far from significance $p = 0.059$). Similarly, the Inter-site variability for the six coastal pilot sites was also significantly different for all grouping categories (i.e. site-type, aspect-type, domain-type). An intra-site variance of attributes for each group of sites out of three groups of site-types (i.e. coastal-waterfront, coastal-waterfront beach, and bayfront) as presented graphically (Figure 4, 5, and 6) varied significantly when analysed for the aspects. However, significant intra-site differences were observed between the sites according to the site-types e.g. coastal-waterfront and coastal waterfront-beach sites, except for the bayfront ($p = 0.622$) (Table 3a). Similarly, significant intra-site differences were observed between the sites for all three site-types and for all the assessed aspects. For the physical domain, attribute scores varied significantly between the sites for the waterfront-beach ($p = 0.028$) and bayfront ($p < 0.001$) site-types but not for the coastal-waterfronts ($p = 0.549$). For the social domain, differences in attribute scores between sites for different site-types were significant except for the waterfront-beach locations ($p = 0.656$). For the aesthetic domain, significant Intra-site variabilities were observed for Bayfront locations only ($p = 0.013$).

For the much larger and more varied Stage-2 sites with their higher inter-rater reliability, the analysis of variance of attribute scores provided a more nuanced pattern (Table 3b). Overall, all 21 Stage-2 sites showed significant inter-site variability for all site-types ($p < 0.001$), all aspect-types ($p < 0.001$) and for all domain-types i.e. physical ($p<0.001$), social ($p<0.001$), and aesthetic ($p<0.001$). To test the intra-site variability of scores between pairs of similar sites, a sub-sample of 18 sites in nine groups were selected (Table 3b). Four out of nine site-types: riverembankments ($p < 0.001$); small lake-park, natural setting ($p<0.001$); dockland-semi natural, built ($p<0.001$); bay-cliff and beach ($p= 0.026$) showed significant intra-site differences when analysed for types of settings, in contrast to six site-types: river-beach, natural ($p<0.001$); river-embankment ($p<0.001$); canal-beach, natural bank ($p<0.001$); fountain-built ($p= 0.007$); coastal-beach ($p= 0.019$); and pond-park ($p<0.001$) when analysed for all aspects (Table 3b).

Out of nine site-types, intra-site differences were statistically significant for the four site-types i.e. river-beach, natural ($p<0.001$); river- embankment walk, natural, built ($p<0.001$); canal-beach, natural bank ($p= 0.001$); coastal- beach ($p= 0.001$)) for the physical domain compared to two site-types i.e. river-beach, natural ($p= 0.017$); and canal-beach, natural bank ($p= 0.021$) for the social domain, and one site-type (i.e. small lake- park, natural setting ($p= 0.030$) for the aesthetic domain). Intra-site differences for attributes between sites in each site-type were found to be non-significant for the social and aesthetic domain, which did not necessarily mean that they were the sites with similar social and aesthetical quality.

Spidergrams were plotted of the physical, social, and aesthetic domains. Each spidergram juxtaposed two sets of scores of the same attributes for two similar sites, allowing for easier visual comparison of the similarities or differences across aspects within or in-between site-types (see Figure 4, 5, and 6 for the six sub-sample pilot sites in three site-types from Stage-1 and Figure 7, 7, and 8 for Stage-2 assessments: the18 paired sub-sample of sites in nine site-type categories). In contrast to the observed significance of intra-site differences for the attributes, for both Stage-1 and Stage-2, all the overlapping spidergrams plotted for each site-type showed visible differences between sites for their attributes scores for different aspects.

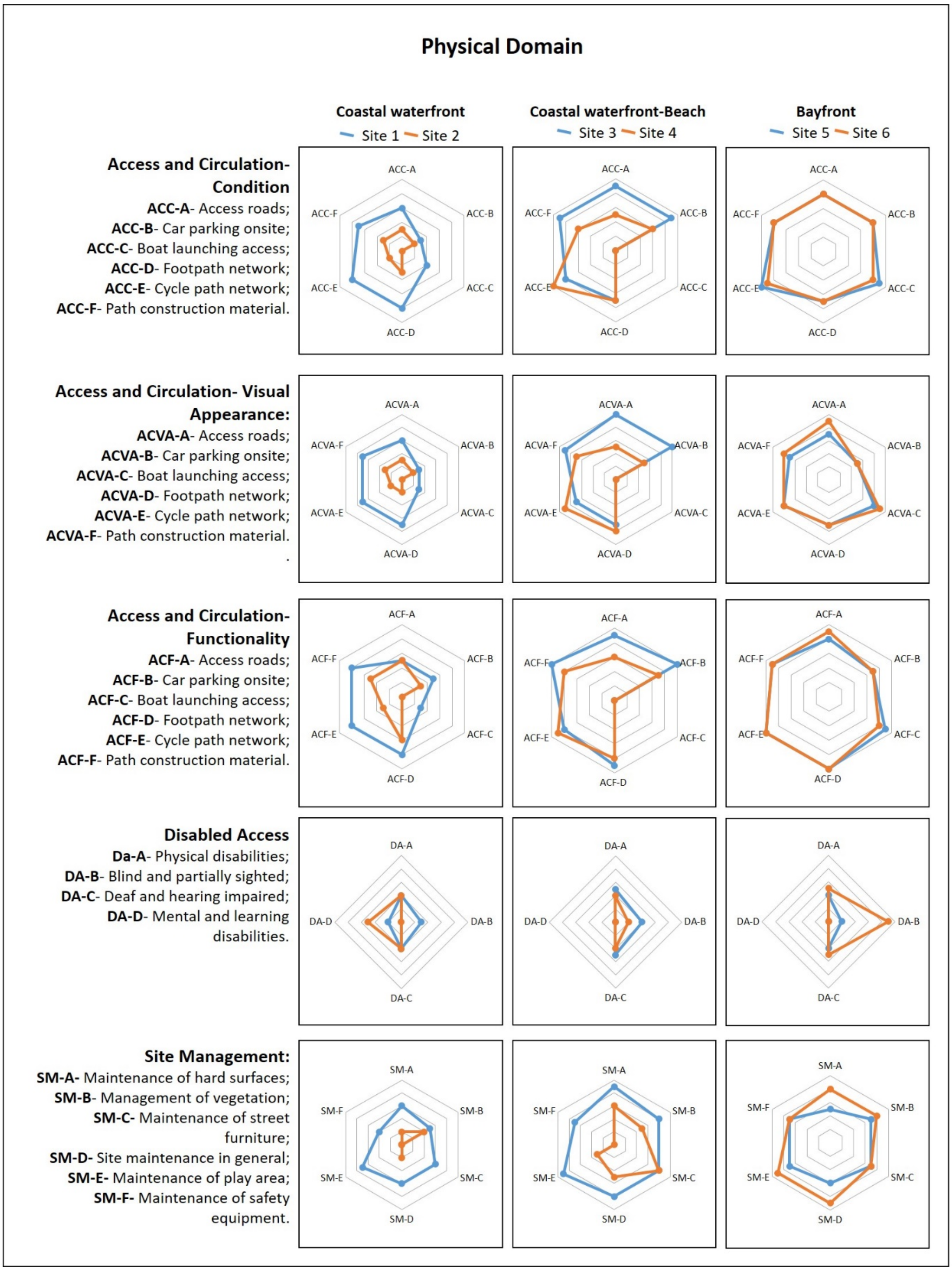

Figure 4: The moderated rating scores for aspects and attributes of the physical domain of the six sub-sample sites of three site types of the Stage-1 pilot test. For all sites the analysis of variance of attribute ratings was significant for the physical domain (p< 0.001).

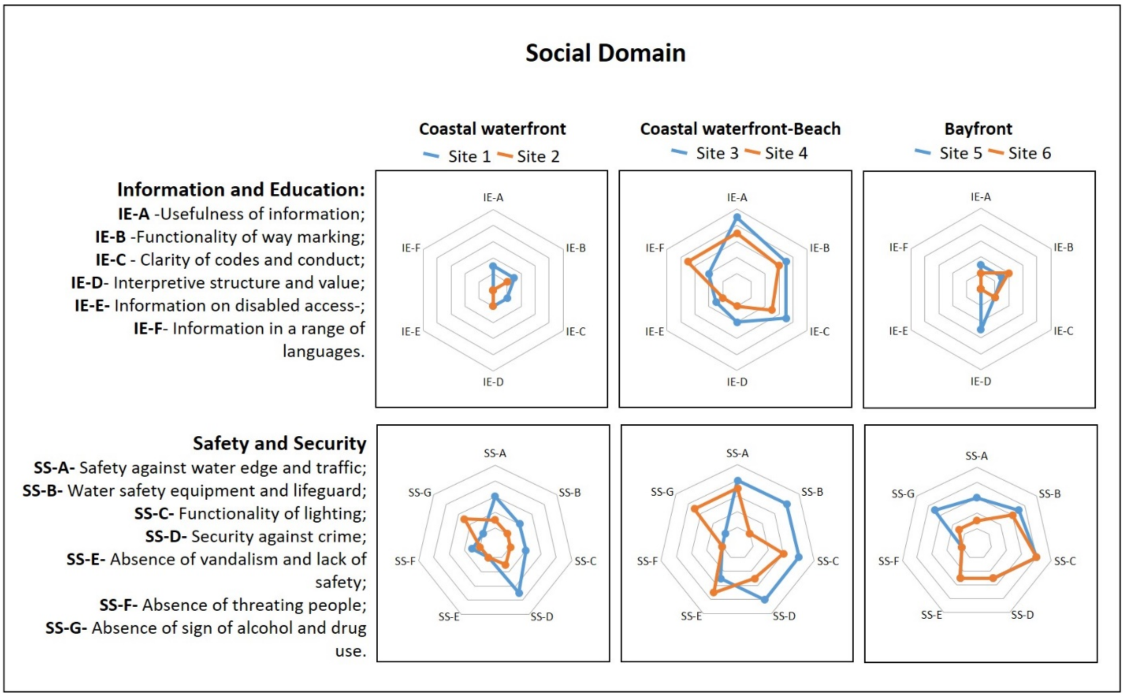

Figure 5: The moderated rating scores for aspects and attributes of the social domain of the six sub-sample sites of three site types of the Stage-1 pilot test. For all sites analysis of variance of attribute ratings was significant for the social domain (p< 0.001).

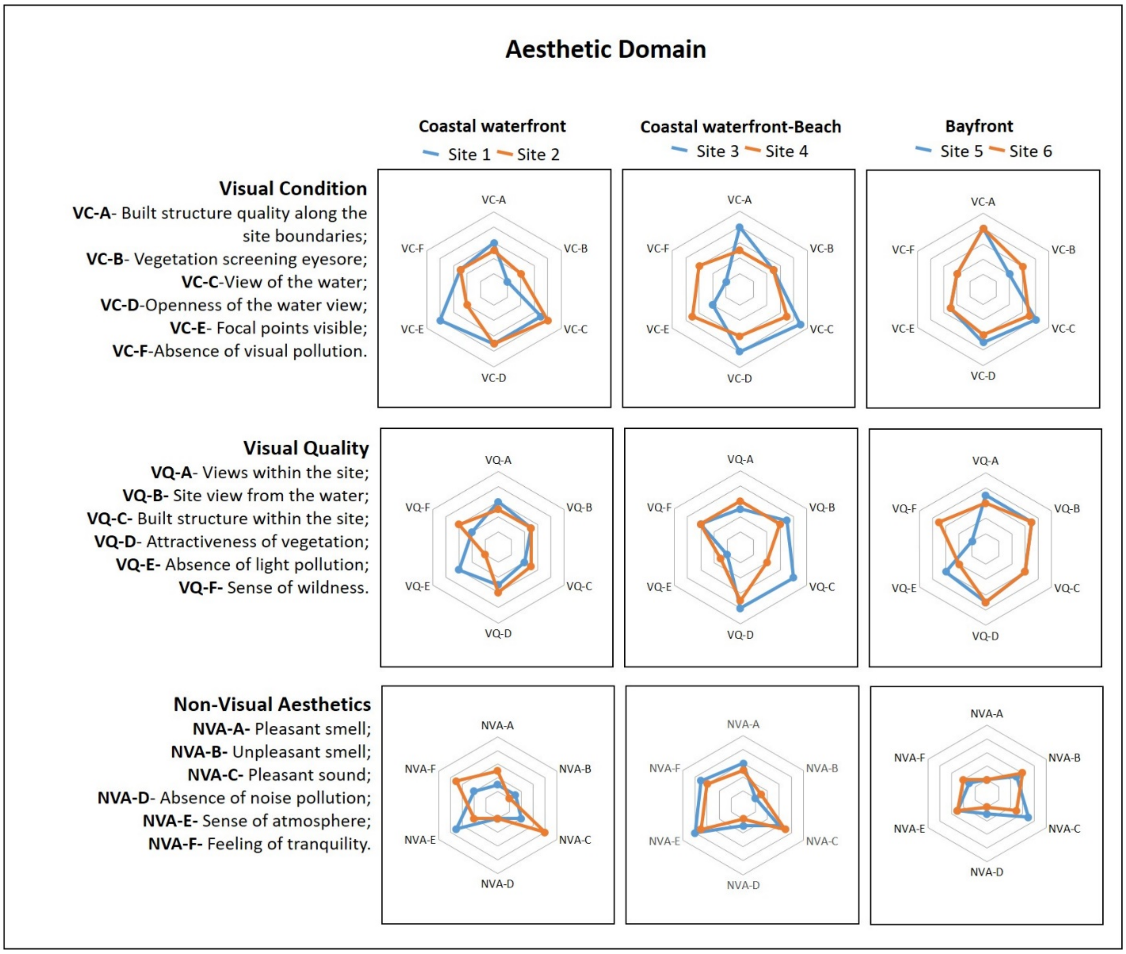

Figure 6: The moderated rating scores for aspects and attributes of the aesthetic of the six sub-sample sites of three site types of the Stage-1 pilot test. For all sites analysis of variance of attribute ratings was significant for the aesthetic domain (p< 0.010).

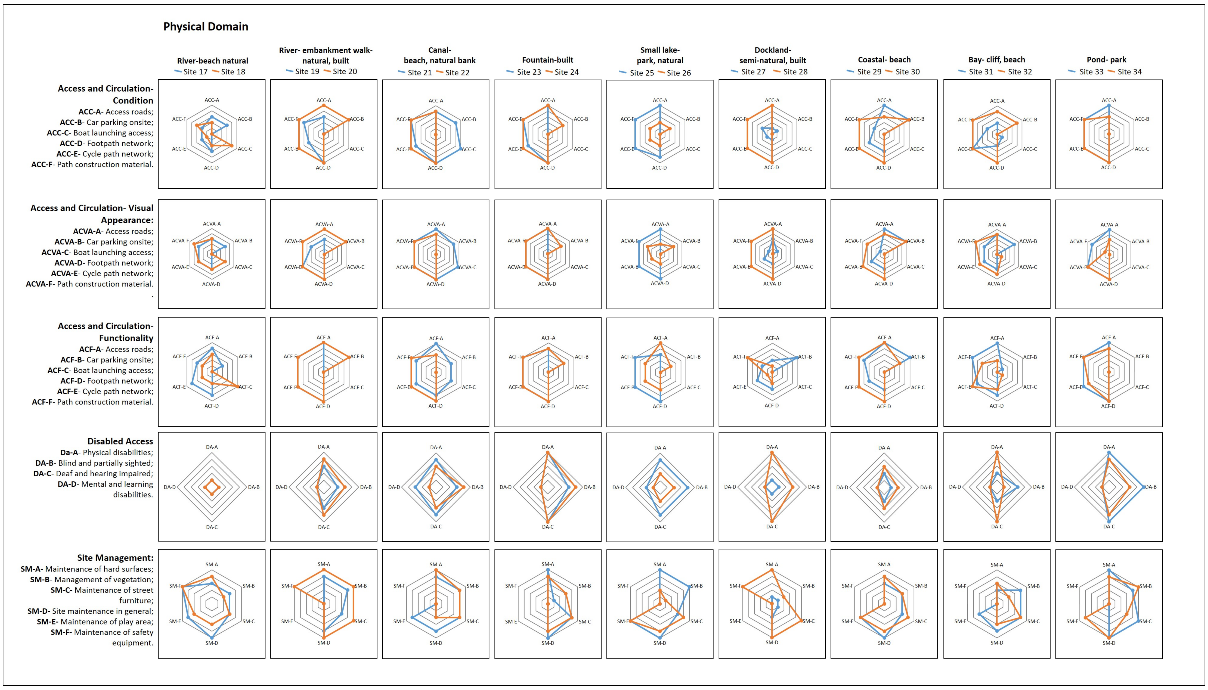

Figure 7: The moderated rating for aspects and attributes of the physical domain of the sub-sample of 18 paired sites of nine site-type of Stage-2 sites. For all sites analysis of variance of attribute ratings was significant for the physical domain (p< 0.001).

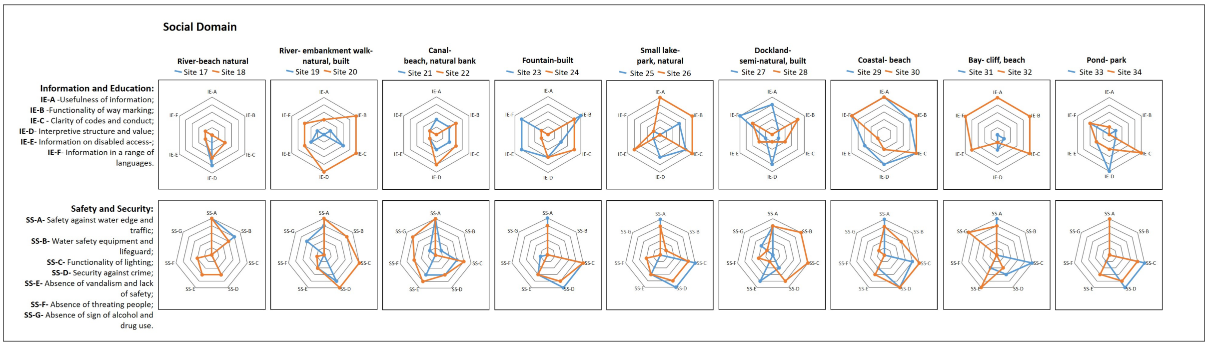

Figure 8: The moderated rating for aspects and attributes of the sub-sample of 18 paired sites of nine site-type of Stage-2 sites . For all sites analysis of variance of attributes ratings was significant for the social domain (p= 0.001).

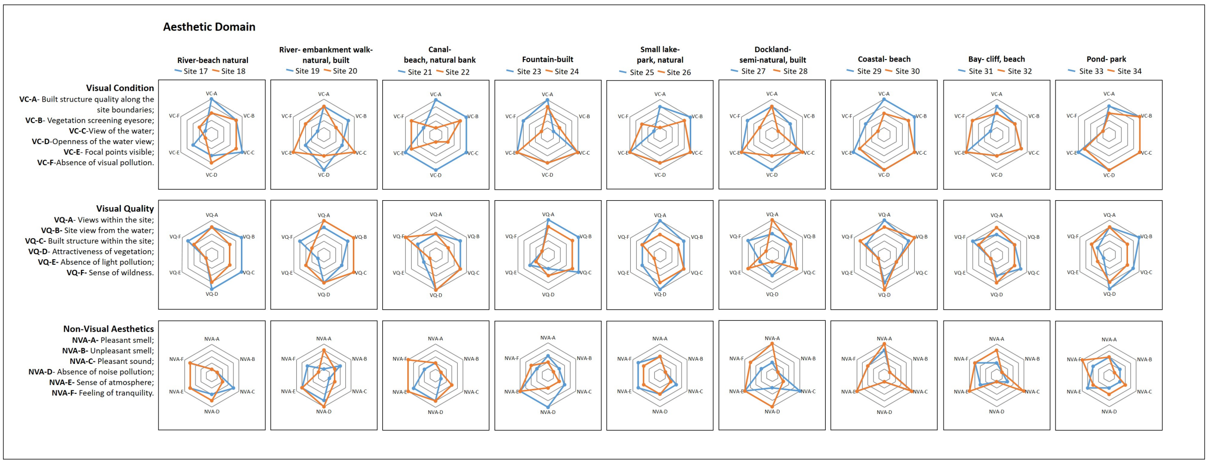

Figure 9: The moderated rating for aspects and attributes of the social domain of the sub-sample of 18 paired sites of nine site-type of Stage-2 sites . For all sites analysis of variance of attribute ratings was significant for the aesthetic domain (p< 0.001).

For the six pilot sites from Stage-1, the attributes for aspects of the physical domain showed large variations for the intra-coastal waterfront sites compared with the intra-coastal waterfront-beach and intra-bayfront sites (for all Stage-1 sites see Supplementary Material Appendix 3, Figure 1). Bayfront sites were selected from within a single geo-climatic and socio-cultural context and displayed similar qualities and characteristics. Conversely, coastal waterfront-beach sites, also selected from within a similar geo-climatic and socio-cultural context, were different from each other in terms of the provision for the access and circulation infrastructure and site management practices. Overall, disabled access conditions were scored low for all sites. For the social domain, large intra-site differences were observed for all site-types, except for bayfront sites. Similarly, for the aesthetic domain, more similarities were observed between sites for the coastal waterfront site-types and bayfront site-types than for the coastal waterfront beach sites. For the attributes for the visual quality aspect, low Intra-site differences were observed between sites of different site-types.

All sites for Stage-2 testing were selected from within a single geo-climatic, and socio-cultural and planning context. However, spidergrams showed large intra-site differences for attributes for all aspects for the physical and social domain. Low intra-site differences were observed for the attributes for aspects in the aesthetic domain. However, spidergrams for all 21 different blue space settings showed inter-site differences for blue space attributes and aspects which represent the qualities and characteristics of the place (see Supplementary Material Appendix 3, Figure 2).

Discussion

Similar to audit instruments developed to assess physical environmental factors that impact physical activities (Joseph and Maddock, 2016; Brownson et al., 2009) there is a growing need for a reliable environmental assessment tool to assess attributes of blue spaces that may impact their design, planning and management, use of the place and physical activities and health-promoting behaviour. It has been established that blue spaces are among people’s favourite outdoor places for recreation (Korpela et al., 2010) and are associated with many benefits such as relaxation and restoration for stress (e.g. White et al., 2013), and these benefits may be linked to different blue space types, qualities and characteristics (White et al., 2010; 2013; Volker and Kistemann,2011). The application of a reliable assessment tool allows for a more systematic evaluation of blue spaces in terms of the relationships within attributes of a single blue space or multiple blue spaces of similar or different types. This may help further to explore the relationships between blue space qualities and to promote evidence-based planning and design of blue spaces for health promotion.

This study tested the inter-rater reliability of the BlueHealth Environmental Assessment Tool (BEAT) and also tested its effectiveness in detecting variations in the health-promoting affordances and in depicting, through the use of spidergrams, the differences in blue space attributes that identify the unique qualities and characteristics of different blue spaces. We observed that in general there is a good degree of reliability – in the 60% plus range for both the pilot Stage-1 and Stage-2 in the un-moderated scores. Common to most tools which use a similar approach to the BEAT (e.g. Gidlow et al., 2012; Saelens et al., 2006; Brownson at al., 2004) the use of two assessors carrying out independent surveys can lead to differences in scoring but as found between pilot Stage-1 and Stage-2, when the same two assessors do the job and when they have received more detailed training and are provided with good, clear guidance, the reliability increases. We can see from the results that some of the attributes within the various BEAT domains and aspects are more open to interpretation or are more difficult or complex to evaluate than others. For example, accessibility to sites by people with different disabilities may need a deeper knowledge of the specific requirements of each disability or impairment type, and the time spent assessing sites may be insufficient to carry out a complete accessibility audit. Equally, some attributes, such as the condition of a building or piece of infrastructure, especially coastal infrastructure, may not be easily determined at a site visit without checking other information. Another aspect which appears to be difficult is that of non-visual aesthetics which, unless both assessors visit the site at the same time, may vary a good deal: traffic noise for example, or the sound of birds.

Overall, when briefly compared to other tools (mostly developed to assess park environments and path characteristics) e.g. of attributes for the aesthetic domain (Brownson et al., 2004; Troped et al., 2006) for which reliability testing has been published, the BEAT performs as well or better, given the comprehensiveness, range of aspects and attributes it covers.

In the pilot testing at Stage-1, the primary assessor was one of the team which developed of the tool and therefore, understandably, knew what was meant by each attribute, while the other assessors were doing it for the first time – and applying something under development with only draft guidance available to read. The probable reason for the lower reliability results for attributes scores, was thus partly due to the mix of local and non-local experts, who had different knowledge about the landscape, site and context, and the level of training they received compared to Stage-2. Therefore, it is no surprise that the reliability was not as good in Stage-1 as in Stage-2. The Stage-1 assessors received much less training (mostly practical aspects of the onsite survey) than those for Stage-2 (in the light of the experiences of Stage-1) and it made a clear difference. For the Stage-2, selection of experts from within a similar socio-cultural and professional background, and the improved training (in the light of the experiences of Stage-1) made a difference as did the deeper acquaintance with the guidance materials.

However, we should not overlook the role of moderation and the discussions necessary to overcome differences – We found in some comments made by the test assessors that this aspect is really useful, especially with regard to the much more subjective attributes such as aesthetic ones. Such discussions can also be of use for subsequent planning – since they identify aspects which need special attention at the planning and design phases (assuming that the BEAT is applied as part of project development). Thus the notes recorded on the forms play an important role in addition to the scores themselves – quantitative and qualitative analysis reveals more than either method used alone.

The strengths of this study include the fact that the BEAT has been developed as a comprehensive tool to assess a wide range of blue space attributes pertaining to the physical, social, and aesthetics domains. By conducting assessments of a wide range of blue space types within different geographic and climatic contexts, we established the fact that the BEAT is a robust, reliable, and effective tool to capture the varying qualities of different blue spaces.

When using scales for scoring attributes there is always a risk that mid-range scores dominate and that a tool such as the BEAT might not be able to pick up real and meaningful differences between different sites. Testing of the effectiveness of the tool using a wide range of both similar and different blue space settings and the analysis of intra- and inter-site variance analysis of the ratings of attributes show that the tool can identify the health-promoting affordances of blue space and that while scores for sites may show similarities in some aspects they show major differences in others – so that the BEAT can be applied with confidence and used as a reliable tool.

Limitations

The testing of the reliability and effectiveness of the BEAT recognised some methodological limitations of the study. There is a notable lack of a variety of blue space samples from diverse geo-political regions of Europe (or elsewhere) which could have improved the robustness of tool. This research used a limited number of sample blue spaces, which may have influenced the effectiveness of the BEAT. We did not use the information collected using BEAT- Steps 1 and 2 (site information at macro and micro level) or qualitative information collected in Step 3; we expect that the full application of the BEAT at each site would have produced a richer result. However, since the aim here was to test the reliability and effectiveness of the quantitative scoring part which gives the tool its robustness, the results demonstrate that we can apply the tool with confidence.

Because we assessed each of the blue spaces at a single point in time, it is also possible that results related to certain characteristics were time- and season-specific, and that repeated assessments of a site during different seasons of the year might be necessary properly to understand a given site.

The use of a common assessor (one of the developers of the tool) at Stage-1 could be seen as a limitation in this study. This assessor assessed all sites for Stage-1 but was unfamiliar with them, he was affiliated to a similar discipline to the other experts, and used the same guidance notes. Moreover, the assessment scores were moderated through discussion before they were analysed, therefore eliminating to some extent possible personal bias in the results. The good inter-rater reliability shows that the tool is robust and that professionals can use produce reliable results, but in planning and design disciplines the score moderation through discussion in relation to the more qualitative data can be expected to provide better results. This best practice was adopted expecting to increase fairness in the assessment process in the absence of a “gold standard” blue space reference site. Nevertheless, a high level of overall agreement between observers in both stages (even is lower in the pilot test) suggests that the BEAT can be applied as a reliable tool in the future.

The BEAT professional version has been designed for professionals and researchers, and following best practice, we adopted a guidance-based assessment approach for which may be seen as a limitation in terms of its usability by non-professional members of local communities. It is probably impossible to achieve 100% reliability in any tool; it is part of the difficulty of assessing sites where not everything can be objectively measured and also where they may be differences in expertise or experience of assessors. A common limitation found in other studies that tested similar environmental audit tools was the reliability of attributes that are open to subjective bias in predicting a particular condition or quality (i.e. visual quality, safety, and disorder, etc.). We identified that in-depth training and guidance with clear explanations could enhance the reliability of such items. This was recognised at the outset of the BEAT development and so the guidance was developed (and has been improved in the light of the attributes for which better descriptions helped to increase reliability).

A final limitation of the BEAT testing could be the use of paper forms as an alternative to the online tool which means it can take more time and this may be a barrier to its uptake. However, in certain circumstances paper forms can be a fall-back (e.g. where wifi is weak or ICT literacy is low) or can allow the assessors to give a more detailed justification for their score.

Conclusions

In this study looking at the results for reliability and effectiveness of the BEAT tool, we had two objectives: to see (a) how the survey tool performs when used by different assessors at different sites, and (b) how effective the tool is at discerning real differences between site quality and characteristics that may inform differences in potential for health promotion.

For objective 1) we found that subject to adequate training and extensive use of the guidance provided, the ratings among different assessors can achieve a close agreement for a majority of aspects – very close for some but with some gaps likely to remain owing to the subjective or complex nature of the attributes being addressed. In Stage-2, where the same two assessors rated the same sites at the same time and also received more training beforehand, the reliability was generally much higher than in Stage-1. For objective 2) we can conclude that the BEAT is effective at identifying and distinguishing between the various aspects of specific sites and that there is no problem of mid-range scores tending to dominate and thus reduce the functionality of the tool.

Declaration of Competing interest

The authors declare no conflicts of interest.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 666773.

References

Arthur, L.M. (1977) Predicting scenic beauty of forest environments: some empirical tests. Forest Science, 23, 151-160.

Ashbullby, K. J., Pahl, S., Webley, P., & White, M. P. (2013). The beach as a setting for families' health promotion: A qualitative study with parents and children living in coastal regions in Southwest England. Health & Place, 23, 138-147.

Bedimo-Rung, A. L., Gustat, J., Tompkins, B. J., Rice, J., & Thomson, J. (2006). Development of a direct observation instrument to measure environmental characteristics of parks for physical activity. Journal of Physical Activity and Health, 3(s1), S176-S189.

Britton, E., Kindermann, G., Domegan, C. and Carlin, C., (2020). Blue care: a systematic review of blue space interventions for health and wellbeing. Health promotion international, 35(1), pp.50-69.

Breen, A., & Rigby, D. (Eds.). (1994). waterfronts: cities reclaim their edge. United State: McGraw-Hill, Inc;

Brownson, R. C., Hoehner, C. M., Brennan, L. K., Cook, R. A., Elliott, M. B., & McMullen, K. M. (2004). Reliability of 2 instruments for auditing the environment for physical activity. Journal of Physical Activity and Health, 1(3), 191-208.

Brown, S., (2020). The social benefits of Blue Space: a systematic review, Environment Agency, Horizon House, Deanery Road, Bristol, ISBN: 978-1-84911-461-5

Brownson, R. C., C. M. Hoehner, K. Day, A. Forsyth, and J. F. Sallis. 2009. "Measuring the Built Environment for Physical Activity. State of the Science." *American Journal of Preventive Medicine* 36(4 SUPPL.): S99-S123.e12.

Cary-Elwes, G. (1996). Urban waterfront regeneration- A precious asset, Landscape Design, No.252, July/August, pp.11-12.

Cavnar, M. M., Kirtland, K. A., Evans, M. H., Wilson, D. K., Williams, J. E., Mixon, G. M., & Henderson, K. A. (2004). Evaluating the Quality of Recreation Facilities: Development of an Assessment Tool. Journal of Park & Recreation Administration, 22(1).

Corbin, A. (1994). The lure of the sea: the discovery of the seaside in the western world, 1750-1840. Univ of California Press.

Dempsey, S., Devine, M. T., Gillespie, T., Lyons, S., & Nolan, A. (2018). Coastal blue space and depression in older adults. Health & place, 54, 110-117.

Depledge, M.H., White, M.P., Maycock, B., & Fleming, L.E. (2019). Time and tide: Our future health and well-being depends on the Oceans. British Medical Journal, 366:l4671.

De Bell, S., Graham, H., Jarvis, S., & White, P. (2017). The importance of nature in mediating social and psychological benefits associated with visits to freshwater blue space. Landscape and Urban Planning, 167, 118-127.

Elliott, L. R., White, M. P., Grellier, J., Rees, S. E., Waters, R. D., & Fleming, L. E. (2018). Recreational visits to marine and coastal environments in England: Where, what, who, why, and when?. Marine Policy, 97, 305-314.

Emery, J., Crump, C., & Bors, P. (2003). Reliability and validity of two instruments designed to assess the walking and bicycling suitability of sidewalks and roads. American Journal of Health Promotion, 18(1), 38-46.

Fang, Y., & Jawitz, J. W. (2019). The evolution of human population distance to water in the USA from 1790 to 2010. Nature communications, 10(1), 1-8.

Feldman M. (1999). Urban Waterfront Regeneration in the Baltic States: The Case of Tallinn. In: Hedegaard L., Lindström B., Joenniemi P., Östhol A., Peschel K., Stålvant CE. (eds) The NEBI Yearbook 1999. Springer, Berlin, Heidelberg

Finlay, J., T. Franke, H. McKay, and J. Sims-Gould. (2015). "Therapeutic Landscapes and Wellbeing in Later Life: Impacts of Blue and Green Spaces for Older Adults." *Health and Place* 34:97–106.

Foley, R., Kistemann, T. (2015). Blue space geographies: Enabling health in place, Health & Place, Volume 35, September 2015, Pages 157-165;

Frumkin, H., G. N. Bratman, S. J. Breslow, B. Cochran, P. H. Kahn, J. J. Lawler, P. S. Levin, P. S. Tandon, U. Varanasi, K. L. Wolf, and S. A. Wood. 2017. "Nature Contact and Human Health: A Research Agenda." *Environ Health Perspect* 125(7):1–18.

Gascon, M., M. Triguero-Mas, D. Martínez, P. Dadvand, J. Forns, A. Plasència, and M. J. Nieuwenhuijsen. 2015. "Mental Health Benefits of Long-Term Exposure to Residential Green and Blue Spaces: A Systematic Review." *Int J Environ Res Public Health* 12(4):4354–79.

Gascon, M., W. Zijlema, C. Vert, M. P. White, and M. J. Nieuwenhuijsen. 2017. "Outdoor Blue Spaces, Human Health and Well-Being: A Systematic Review of Quantitative Studies." *Int J Hyg Envir Heal* 4(8).

Garrett, J., White, M. P., Huang, J., Ng, S., Hui, Z., Leung, C., ... & Depledge, M. H. (2019). The association between blue space exposure, health and well-being in Hong Kong. Health Place, 55, 100-110..

Gidlow, C. J., N. J. Ellis, and S. Bostock. 2012. "Development of the Neighbourhood Green Space Tool (NGST)." *Landscape and Urban Planning* 106(4):347–58.

Grellier, J., M. P. White, M. Albin, S. Bell, L. R. Elliott, M. Gascon, S. Gualdi, L. Mancini, M. J. Nieuwenhuijsen, D. A. Sarigiannis, M. van den Bosch, T. Wolf, S. Wuijts, and L. E. Fleming. 2017. "BlueHealth: A Study Programme Protocol for Mapping and Quantifying the Potential Benefits to Public Health and Well-Being from Europe's Blue Spaces." *BMJ Open* 7(e016188):1–10.

Giraudeau, B. (1996). Negative values of the intraclass correlation coefficient are not theoretically possible. Journal of clinical epidemiology, 49(10), 1205.

Haeffner, M., Jackson-Smith, D., Buchert, M., & Risley, J. (2017). Accessing blue spaces: social and geographic factors structuring familiarity with, use of, and appreciation of urban waterways. Landscape and Urban Planning, 167, 136-146.

Hartig, T., R. Mitchell, S. de Vries, and H. Frumkin. 2014. "Nature and Health." *Annual Review of Public Health* 35(December 2013):207–28.

Hoyle, B., & Pinder, D. (1981). City port industrialization and regional development: Oxford: Pergamon Press;

Hoyle, B., & Pinder, D. (Eds.). (1992). European port cities in transition (1st edition ed.). London: Belhaven Press;

Hoyle, B. (2002). Urban waterfront revitalization in developing countries: the example of Zanzibar's Stone Town. The Geographical Journal, 168(2), 141-162;

IBM Corp. Released in 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp

Joseph, R. P., & Maddock, J. E. (2016). Peer-Reviewed: Comparative Analysis of Five Observational Audit Tools to Assess the Physical Environment of Parks for Physical Activity, 2016. Preventing chronic disease, 13.

Kabisch, N., van den Bosch, M., & Lafortezza, R. (2017). The health benefits of nature-based solutions to urbanization challenges for children and the elderly–A systematic review. Environmental Research, 159, 362-373.

Korpela, Kalevi M., Matti Ylén, Liisa Tyrväinen, and Harri Silvennoinen. 2010. "Favorite Green, Waterside and Urban Environments, Restorative Experiences, and Perceived Health in Finland." *Health Promotion International* 25(2):200–209.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). IBM SPSS for intermediate statistics: Use and interpretation. Routledge.

Liljequist D, Elfving B, Skavberg Roaldsen K (2019) Intraclass correlation – A discussion and demonstration of basic features. PLoS ONE 14(7): e0219854. https://doi.org/10.1371/journal.pone.0219854

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A. M., ... & Lupp, G. (2017). Exploring pathways linking greenspace to health: theoretical and methodological guidance. Environmental Research, 158, 301-317.

McKight, P. E. and J. Najab. 2010. “The Kruskal-Wallis Test.” in The Corsini Encyclopedia of Psychology.

Miller, D., J. Roe, C. Brown, S. Morris, J. Morrice, and C. Ward Thompson. 2012. *Blue Health: Water, Health, and Well-Being*. Aberdeen.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. Annals of cardiac anaesthesia, 22(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Mishra, H. S., S. Bell, P. Vassiljev, F. Kuhlmann, G. Niin, and J. Grellier. 2020. “The Development of a Tool for Assessing the Environmental Qualities of Urban Blue Spaces.” *Urban Forestry & Urban Greening* 49(January):126575.

Nasar, J. L., & Li, M. (2004). Landscape mirror: the attractiveness of reflecting water. Landscape and Urban Planning, 66(4), 233-238.

Nutsford, D., Pearson, A. L., Kingham, S., & Reitsma, F. (2016). Residential exposure to visible blue space (but not green space) associated with lower psychological distress in a capital city. Health & Place, 39, 70-78.

Ott, I; Laarmaa, R; Saar, K; Mishra, H S; Drill, M; Vasilljev, P; Balicka, J; Kuhlmann, F; Niin, G; and Bell, S, (2017). Designing of methods for sustainable functioning of Blue-space areas in Europe. NBS2017: Nature-based solutions - From innovation to common-use; 24-26.10.2017; Tallinn. Tallinn: Tallinn University, 107.

Papathanasopoulou, E., White, M.P., Hattam, C., Lannin, A., Harvey, A., & Spencer, A. (2016). Valuing the health benefits of physical activities in the marine environment and their importance for marine spatial planning. Marine Policy, 63, 144-152.

Pasanen, T., White M.P., Wheeler, B., Garrett, J., Elliott, L. (2019). Neighbourhood blue space, health and wellbeing: The mediating role of different types of physical activity. Environment International, 131, 105136.

Raymond, C. M., Gottwald, S., Kuoppa, J., & Kyttae, M. (2016). Integrating multiple elements of environmental justice into urban blue space planning using public participation geographic information systems. Landscape and Urban Planning, 153, 198-208.

Saelens, B. E., Frank, L. D., Auffrey, C., Whitaker, R. C., Burdette, H. L., & Colabianchi, N. (2006). Measuring physical environments of parks and playgrounds: EAPRS instrument development and inter-rater reliability. Journal of Physical Activity and Health, 3(s1), S190-S207.

Suresh, M., Smith, D. J., & Franz, J. M. (2006). Person environment relationships to health and wellbeing: An integrated approach. idea journal, 2006, 87-102.

Solomon, S. (2010). Water: The Epic Struggle for Wealth. Power and Civilization, New York: Harper. available online at https://publicism.info/environment/water/index.html, accessed on 15.11.2020.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273-1296.

Triguero-Mas, M., Dadvand, P., Cirach, M., Martínez, D., Medina, A., Mompart, A., ... & Nieuwenhuijsen, M. J. (2015). Natural outdoor environments and mental and physical health: relationships and mechanisms. Environment international, 77, 35-41.

Troped, P. J., Cromley, E. K., Fragala, M. S., Melly, S. J., Hasbrouck, H. H., Gortmaker, S. L., & Brownson, R. C. (2006). Development and reliability and validity testing of an audit tool for trail/path characteristics: the Path Environment Audit Tool (PEAT). Journal of Physical Activity and Health, 3(s1), S158-S175.

Völker, S. and T. Kistemann. 2011. "The Impact of Blue Space on Human Health and Well-Being - Salutogenetic Health Effects of Inland Surface Waters: A Review." *International Journal of Hygiene and Environmental Health* 214(6):449–60.

Völker, S. and T. Kistemann. 2013. "Reprint of: 'I'm Always Entirely Happy When I'm Here!' Urban Blue Enhancing Human Health and Well-Being in Cologne and Düsseldorf, Germany." *Social Science and Medicine* 91:141–52.

Völker, S. and T. Kistemann. 2015. "Developing the Urban Blue: Comparative Health Responses to Blue and Green Urban Open Spaces in Germany." *Health & Place* 35:196–205.

Völker, S., J. Matros, and T. Claßen. 2016. "Determining Urban Open Spaces for Health-Related Appropriations: A Qualitative Analysis on the Significance of Blue Space." *Environ Earth Sciences* 75(13).

Völker, & Kistemann, (2015). Developing the urban blue: comparative health responses to blue and green urban open spaces in Germany. Health & Place, 35, 196-205.

Walton, J. K. (1983) The English Seaside Resort: a Social History 1750–1914

Wheeler, B. W., Rebecca Lovell, Sahran L. Higgins, M. P. White, I. Alcock, Nicholas J. Osborne, Kerryn Husk, Clive E. Sabel, and M. H. Depledge. 2015. “Beyond Greenspace: An Ecological Study of Population General Health and Indicators of Natural Environment Type and Quality.” *Int J Health Geogr* 14:17.

Wheeler, B. W., M. P. White, Will Stahl-Timmins, and M. H. Depledge. 2012. “Does Living by the Coast Improve Health and Wellbeing.” *Health and Place* 18(5):1198–1201.

White, M. P., A. Smith, K. Humphryes, S. Pahl, D. Snelling, and M. H. Depledge. (2010). “Blue Space: The Importance of Water for Preference, Affect, and Restorativeness Ratings of Natural and Built Scenes.” *Journal of Environmental Psychology* 30(4):482–93.

White, M.P., Pahl, S. Ashbullby, K.J., Herbert, S.& Depledge, M.H. (2013). Feelings of restoration from recent nature visits. Journal of Environmental Psychology, 35, 40-51.

White, M.P., Wheeler, B.W., Herbert, S., Alcock, I. & Depledge, M.H. (2014). Coastal proximity and physical activity. Is the coast an underappreciated public health resource? Preventive Medicine, 69, 135-140.categorical data. *Biometrics*, *33*, 159-174.

White, M. P., Deborah Cracknell, Abigail Corcoran, Gemma Jenkinson, and M. H. Depledge. 2014. “Do Preferences for Waterscapes Persist in Inclement Weather and Extend to Sub-Aquatic Scenes?” *Landscape Research* 39(4):339–58.

White, M.P., Pahl, S., Wheeler, B.W., Fleming, L.E.F., & Depledge, M.H. (2016). The ‘Blue Gym’: What can blue space do for you and what can you do for blue space? Journal of the Marine Biological Association,96, 5-12.

White, M. P., Elliott, L. R., Gascon, M., Roberts, B., & Fleming, L. E. (2020). Blue space, health and well-being: A narrative overview and synthesis of potential benefits. Environmental Research, 110169.

World Health Organisation (2014). Global report on drowning: Preventing a leading killer. Downloaded 22nd Jan 2020 from: http://www.who.int/violence_injury_prevention/publications/drowning_global_report/Final_report_full_web.pdf

Willes, K.L. (2018) Data Cleaning In: Allen, M. (Ed.). (2017). The SAGE Encyclopedia of Communication Research Methods. Sage Publications.

Yevjevich, V. (1992) Water and Civilization, Water International, 17:4, 163-171, DOI: 10.1080/02508069208686135

Zhang, Y., A. E. Van den Berg, T. Van Dijk, and G. Weitkamp. 2017. “Quality over Quantity: Contribution of Urban Green Space to Neighborhood Satisfaction.” *Int J Environ Res Public Health* 14(5).