UCL DEPARTMENT OF GEOGRAPHY

Assessing the reliability of the Anglers' Riverfly Monitoring Initiative (ARMI)

Caroline Cahill

10,375 words

Thesis submitted for the degree of MSc Aquatic Science, Dept of Geography,
UCL (University College London)

Supervisor: Steve Brooks

September 2019

Abstract

1. The Angler's Riverfly Monitoring Initiative (ARMI) is a highly active citizen science scheme in the UK, in which trained volunteers gather river water quality data every month by collecting, sorting and scoring macroinvertebrate samples. Ensuring the collected data is of a high quality is vital so the data can be fully utilised by regulatory authorities. However, no analysis so far has considered how ARMI sample scores may vary between volunteers, and the reasons for such variation, knowledge of which is essential to confirm and improve the reliability of the scheme. It is also unknown if there are minimum and maximum sample sorting time recommendations.

2. To assess how sample scores vary between volunteers, the overall and taxa-specific ARMI scores gathered over three months by 13 volunteer groups at their regular sites were compared with scores generated at the same site and time by the main investigator, also known as assessing inter-rater reliability. To understand the causes of low inter-rater reliability between volunteers and the main investigator, the influence of site water quality and inter-sampler differences (specifically, differences in: method used to complete the kick sample and stone search, group size, and sample sorting time) on the score differences were analysed.

3. Overall, ARM achieved 'good' inter-rater reliability, and neither site water quality nor inter-sampler differences were found to influence overall score differences between the main investigator and volunteers. However, this was not the case taxa-specifically, with cased caddisfly, Baetdiae, Ephemerellidae and caseless caddisfly achieving only 'moderate- poor' inter-rater reliability, and score differences for cased caddisfly found to be influenced by differences in kick sample method and group size, and Ephemerellidae score differences found to be influenced by differences in group size and sample sorting time differences. There was no time point during sample sorting at which score gains became more or less likely.

4. These findings suggest that, overall, ARMI is a reliable scheme. Hence, reports of low water quality should be acted on by regulatory authorities, datasets should be analysed to understand temporal and spatial trends in water quality, and regulatory authorities should continue with supporting, funding and expanding the scheme. Strategies are suggested to increase the likelihood that volunteers will use the standard kick sample method, which should help improve the low inter-rater reliability of cased caddisfly. It is also recommended that unless unavoidable, volunteer monitoring units should be composed of pairs rather than individuals, which would hopefully improve the reliability of both Ephemerellidae and cased caddisfly scores. It is also recommended that volunteers spend as long as they feel appropriate analysing their sample.

Word count: 10,375

Acknowledgements

Thanks go to Steve Brooks for his guidance and support with planning, carrying out and writing up the project, and to Jan Axmacher for his help with the statistical analyses. Finally, thanks to the volunteers for accommodating me during their monitoring and providing much need transportation.

Contents

1. Introduction

Citizen science is the practice of involving non-expert volunteers in scientific research (Carr 2004), with roles ranging from collecting, analysing and disseminating data, to even designing research projects with scientists (Roy et al. 2012). Often developed in response to public concern about anthropogenic impacts on the environment (Conrad & Daoust 2008), and concern about government monitoring of ecosystems (Pollock & Whitelaw 2005), citizen science schemes are utilised to gather large quantities of data (Dickinson et al. 2012) and as a way of engaging and educating citizens (Dickinson et al. 2010) in a variety of ecological and environmental research fields. These include climate change, invasive species, conservation biology, population ecology and water quality monitoring (Silvertown 2009).

Volunteer monitoring programmes have been specifically used to monitor water quality and inform management since the founding of the Izaak Walton League of America (IWLA), an environmental organisation responsible for the launch of an initiative to observe and report observable problems with water quality and pollution in the United States nearly 100 years ago (Firehock & West 1995). Throughout the 20th century, volunteer monitoring techniques and tools developed in sophistication, and by the early 1970's had progressed from measuring water quality with simple observations of litter or strange colours or odours as used by the IWLA, to taking chemical and biological measurements. This occurred both in the United States, such as the Save Our Streams and Stream Quality Monitoring (Firehock & West 1995) surveys, as well as in the United Kingdom, for example the Advisory Centre for Education (Mellanby 1974) and Riverwatch (National Riverwatch 1994) river water quality surveys.

Since then, the field of water quality citizen science has continued to expand, and in recent years a number of citizen science projects have been successfully introduced or trialled to collect water quality data around the world (Thornhill et al. 2019). This is not surprising given the extent of rivers polluted worldwide (EEA 2015; UNEP 2016; Commission for Environmental Cooperation Undated) and the reliance of local communities on healthy freshwater for various reasons, including the provision of clean drinking water, sustaining agricultural production, supporting biodiversity, and its amenity values (UN Water 2016). Many of the current citizen science programmes are centred around directly measuring specific pollutant levels, including nutrients (Castilla et al. 2015; Thornhill et al. 2018; Weigelhofer et al. 2019), organic and inorganic pollution (Levesque et al. 2017; Miguel-Chinchilla et al. 2019) litter and microplastic pollution (Forrest et al. 2019; Kiessling et al. 2019), groundwater pollution (Dawson et al. 2019) and metal contamination (Turner et al. 2017).

Several other schemes, however, are simplified versions of professional biotic indices, and assess the macroinvertebrate taxa present in the water against their known sensitivity/tolerance to different levels of pollution (Kripa et al. 2012), detecting the incidence of pollution through its ecological impacts as opposed to the pollutant itself. Using macroinvertebrates is the most common way of professionally assessing water quality in Europe (Abassi & Abassi 2012; Ramos-Merchante & Prenda 2017), and is also widely used throughout the rest of the world. This is because benthic macroinvertebrates are made

up of species from a wide range of trophic levels and pollution tolerances, reflect prevailing conditions in a foreseeable way, and respond quickly to environmental stress (Metcalfe 1989; Cairns & Pratt 1993; Li et al. 2010; Kripa et al. 2012). Various macroinvertebrate water quality citizen science programmes have been successfully implemented worldwide, including in both lentic (Latimore & Steen 2014; Rose et al. 2016) and lotic (Latimore & Steen 2014; Moffett & Neale 2015; Edwards 2016; Storey et al. 2016; Brooks et al. 2019; Franca et al. 2019) freshwaters.

In the UK, the primary nationwide water quality citizen science scheme for rivers is the Anglers' Riverfly Monitoring Initiative, also known as ARMI. ARMI has been underway since 2007, and is currently active in 35 regional hubs around the UK, on 1850 sites (Figure 1), with 2600 trained ARMI volunteers, many of whom are anglers (Environment Agency 2019). ARMI has had much success so far, detecting numerous pollution incidents around the country (Environment Agency 2019; Fitch et al. 2018), and creating a sizeable database of water quality records, at a large temporal and spatial extent (Brooks et al. 2019), with wider benefits for public awareness of environmental issues, public engagement with science and policy, and public well-being (Brooks et al. 2019; Dunkley 2019).

However, there are also various challenges that citizen science programmes must overcome to be successful, including organisational, data quantity and quality, and data use, issues (Conrad & Hilchey 2011). For ARMI, a significant focus has been ensuring a high quantity and quality of data is collected (Brooks et al. 2019). Progress has been made in this area by ensuring that data is collected at a large extent, using a simplified but recognised method, and that ARMI data is comparable with professional data (Brooks et al. 2019). However, the variation in results obtained by different samplers, and how site water quality and inter-sampler differences might influence the scores, has not yet been investigated (Brooks et al. 2019). Such an analysis is important so that, if necessary, the impacts of any inter-sampler and site water quality differences can be accounted for during analysis and use of the ARMI data (Cooper et al. 2012), as well as to fully confirm how reliable the scheme is, and indicate whether and what improvements could be made to the protocol or training to increase reliability (Bonney et al. 2014; Ramos-Merchante & Prenda 2017; Tredick et al. 2017). Achieving a high level of reliability is important so that regulatory authorities trust the data and fully engage with it and act on it (Tredick et al. 2017). Use of ARMI data by regulatory authorities has the potential to greatly progress river management (Edwards 2016), whilst simultaneously improving volunteer recruitment and retention, as well as funding opportunities, as the scheme and data are demonstrated to be impactful (Thornhill et al. 2019).

1.1 The Anglers' Riverfly Monitoring Initiative

Based on information provided by Brooks et al. (2019) unless otherwise stated, ARMI is a collaborative type of citizen science project (Thornhill et al. 2019), developed together by the Environment Agency (EA) and the Riverfly Partnership (RP). The Riverfly Partnership is partnership of organisations and individuals from across the UK, including anglers, conservationists, scientists, water course managers

Fig 1. ARMI UK Site Network Map; taken from Fitch et al. (2018).

and regulatory authorities, working together to protect and improve the health and quality of rivers, and conserve riverflies (in particular Trichoptera, Ephemeroptera and Plecoperta) and their habitats. At the time of the inception of the RP, and later the ARMI, there was large concern among citizen stakeholders about declines in river quality, and perceptions that the regulatory authorities were not adequately detecting and remediating this decline due to limitations in resources and funding. The ARMI was therefore established to provide individuals and local communities a formal structure with which they

themselves can use standardised, recognised methods to undertake regular reliable water quality assessments on their local river, and communicate and discuss the findings of these assessments with the regulatory authorities. This ensures that incidences of low water quality are being detected and remediated, and high water quality standards are being upheld elsewhere. The scheme is currently hosted by the Freshwater Biological Association, and training of volunteers and operation of the scheme is financed via matched funding; ARMI groups raise money locally, including from local government, water companies and conservation trusts, as well as National Lottery grants, which is matched by the EA (Environment Agency 2019). Between 2010-2016, for example, £250,000 of matched funding was raised for ARMI work (Environment Agency 2016).

The ARMI protocol is a simplified version of the Biological Monitoring Working Party (BMWP), a biotic index widely used in the UK (Di Fiore & Fitch 2016). It works on the premise that different taxa have different levels of sensitivity/ tolerance to organic pollution (Paisley et al. 2014), and therefore the presence and abundance of particular taxa reflects distinct pollution levels (Di Fiore & Fitch 2016). The ARMI protocol consists of a three-minute kick sample, followed by a one-minute manual search of large, liftable stones. Sub-samples are then transferred into the large sorting tray, from which relevant taxa (cased caddis, caseless caddis, Ephemeridae, Ephemerellidae, Heptageniidae, Baetidae, Plecoptera and Gammarus) are picked out live and sorted according to taxon in the segmented tray. These taxa were chosen due to their distribution in rivers around the UK, year-round presence (except for Ephemerellidae), and because they are familiar to most anglers and easy to identify at this resolution (Di Fiore & Fitch 2016). The abundance of these invertebrates are counted/estimated, and awarded a score, according to Table 1. The score for each taxon is added to give an overall score for the sample, with higher scores indicating sites have better water quality. If there are no individuals, or no live in individuals for a certain taxon, that taxon achieves a score of 0.

Table 1. Logarithmic scoring system used for each taxon for each sample

Abundance	Score	Estimated Number
1-9	1	Quick count
10-99	2	Nearest 10
100-999	3	Nearest 100
1000+	4	Nearest 1000

Volunteers are trained in using the simplified protocol and standard equipment during a 1-day workshop. This is led by an RP-accredited trainer, although the local regulatory authority officer often also attends to assist, sometimes alongside current ARMI volunteers. At the training day, volunteers are provided with a laminated fold-out sheet, produced by the Field Studies Council (FSC), called the ‘Riverfly Monitoring Guide’. This fold-out gives an identification guide to the eight taxa of interest (specifically, high resolution images labelled with morphological features of interest), guidance on the sampling procedure and scoring system, advice on health and safety and biosecurity, and instructions

on how to report their scores. In the morning session of the training, the RP-accredited trainer leads the volunteers through a standard presentation, which provides an introduction to the basis of biotic water quality monitoring and the ARMI, explains the Riverfly Monitoring Guide and gives further identification advice on the eight target taxa, including videos showing their movement. In the afternoon session of the day, volunteers are then given a practical demonstration of how to undertake a sample, identify the taxa and estimate their abundance, with an opportunity for them to subsequently practice the collection and sample sorting procedure themselves in groups. Volunteers are helped to calculate a score based on their sample, and are also shown where to report their score on the Riverfly website. A rounding-up session then recaps what the volunteers have learnt during the training day, and advises volunteers on their next steps regarding finding a site to monitor and obtaining equipment (The Riverfly Partnership 2017).

Volunteers often choose to monitor a site in their local area, and normally one which the regulatory authority is not currently monitoring. The standard is to then monitor once per month, every month of the year. Based on the long-term ARMI monitoring data for the site, a trigger level much below the usual score for the site is set by the regulatory authority. If, during their monthly monitoring, a volunteer finds the score is at or below the level, a serious pollution event may have occurred. Volunteers repeat the sample to confirm the trigger level breach, and inform their local coordinator who will contact the regulatory authority. The regulatory authority then attends the site, further investigates and arranges for appropriate remediation of the cause of the poor water quality.

1.2 Importance of ARMI

ARMI is highly active and popular across the country because it brings with it various benefits, for the volunteers, regulatory organisations and the river ecosystems themselves.

Complementing routine monitoring

The Environment Agency are responsible for monitoring river water quality across the UK, and for detecting and arranging for the remediation of pollution (Environment Agency 2018). However, substantial cuts in the budgets of regulatory authorities over recent years have reduced the frequency and number of sites they can monitor water quality (Boren & Scott 2018). With sites located across the UK, and monitoring taking place on a monthly basis, ARMI has the potential to greatly complement the routine statutory monitoring and provide an early warning to pollution by collecting data at a larger spatial and temporal extent than regulatory authorities currently can, in a cost-effective manner (Fitch et al. 2018), with a recent study showing that ARMI data is comparable to professional data (Brooks et al. 2019). This has various benefits, including that it raises the likelihood of detecting incidences of low water quality, and their causes, so that appropriate remediation can be undertaken as soon as possible

after the pollution event. In 2017, for example, 235 confirmed trigger level breaches around the UK were detected first by ARMI volunteers (Fitch et al. 2018), with numerous case studies from throughout ARMI's history also testament to the success of this approach.

Organic pollution, for instance, is one of the most common causes of low water quality detected first by volunteers, which is unsurprising given that the taxa used in the ARMI were chosen for their sensitivity to organic pollution (Di Fiore and Fitch 2016). As reported by Brooks et al. (2019), for example, between September 2015-2016 on Broughton Beck, North Yorkshire, declines in ARMI scores resulted in the EA discovering that a sewage treatment works was discharging poor quality water, which the sewage treatment company are currently addressing at the request of the EA. Similarly, in 2017, on Spratford Stream on the River Culm, Devon, ARMI sampling highlighted low invertebrate diversity, with large numbers of Gammarus but very little else present. Investigatory work by the EA confirmed the low biodiversity, and it was suggested that agricultural organic pollution is likely responsible, with further investigatory work still underway (Environment Agency 2019). In both cases, these issues would not have been detected, or detected so soon, by relying alone on the statutory monitoring carried out by the EA (Environment Agency 2019).

In various instances ARMI scores have also been successfully used to detect other causes of low water quality asides from organic pollution, however. These include low flows, siltation, fish poaching, slurry pollution, road-run off pollution, metal pollution, pesticide pollution, changed river morphology, glycol entering rivers, and leaking sewage on urban rivers (Fitch et al. 2018). For example, on the River Crane, London, between 2014-2017 over 10 incidences of trigger level breaches were reported, with the cause traced to high levels of phosphates and ammonia originating from domestic misconnections via outfalls, as well as cross-connections between foul and surface water sewerage systems. As a consequence, further monitoring and remediation work is now underway by both volunteers and Thames Water to locate and reduce these connections (Brooks et al. 2019). Similarly, in March 2010 on the River Derwent, Northumberland, ARMI volunteers carried out additional samples at the request of the EA over concerns that high levels of rainfall may have washed heavy metals from the old mine workings into the river (Brooks et al. 2019). The scores for late March were lower than for before the high rainfall, reflecting the possible pollution, but they did not breach trigger levels, and the following month returned to normal. In this instance the ARMI scores were useful for the EA and local communities as they advised that while some pollution had occurred, the EA did not need to use important resources for remediation. A point-source pesticide spill was also first noticed by ARMI volunteers on the River Kennett in July 2013, which the EA were able to quickly resolve once notified (Thompson et al. 2016).

Pollution initially detected by ARMI volunteers has even resulted in prosecutions against polluters. In April 2007, for instance, a member of The Rhymney and Sirhowy Flylife monitoring group reported dead fish in the River Sirhowy, South Wales (Riverfly Partnership 2007). Upon investigation by the Environment Agency Wales (EAW), a company were found to have allowed caustic and highly contaminated wastewater from their tanks and treatment plant to run untreated into a surface water

drain that flowed directly into the River Sirhowy. The case was taken to court by the EAW, where the polluters pleaded guilty to two charges, and were fined £4450. The presence of regular, dedicated ARMI monitors on a river and the threat of legal action can therefore serve as a deterrent to potential polluters.

Once remediation strategies have been put in place, the continual and long-term nature of ARMI data is also useful for helping to determine how effective remediation strategies have been, and whether rivers have recovered to pre-pollution standards (Bartle 2018). Furthermore, even if no pollution is detected, the long-term datasets gathered by volunteers are still useful to generate an evidence base to help with resolving pollution investigations that may occur in the future, and inform statutory agencies that despite their own infrequent monitoring, staff resources are not needed on sites which attain high scores (Brooks et al. 2019). However, it is not just incidences of low water quality that ARMI monitoring is useful for. Adopting ARMI monitoring for use in a before-after-control-impact (BACI) set up was used to demonstrate the immediate and long-term positive impact of weir removal on invertebrate communities in the River Bulbourne in Boxmoor (Brooks et al. 2019). ARMI volunteers can also play an important role in the detection and reporting of invasive 'Alert' species, such as the killer shrimp (*Dikerogammarus villosus*), and rare species, such as the Yellow Mayfly (*Potamanthus luteus*) (Fitch 2017).

In complementing statutory monitoring in this way, ARMI provides regulatory authorities with significant financial in-kind benefits (estimated to be worth at least £608,975 for the Environment Agency between 2016-2017) (Fitch et al. 2018). This allows the regulatory authorities to focus their own expert resources on arranging and enforcing the remediation, as well as prosecuting polluters; tasks that cannot be easily nor effectively undertaken by volunteers alone. This ensures that all aspects of protecting and improving river water quality can be achieved.

Increasing environmental science democratisation and citizen engagement with environmental issues

Increased environmental science democratisation, the process of making environmental science and expertise more accessible to the public, whilst at the same time making local knowledge and expertise more accessible to scientists (Carolan 2006), is also achieved through ARMI. Often, this leads to increased citizen engagement with environmental issues, all of which have many benefits for the river ecosystem.

For instance, as discussed previously, ARMI offers regulatory authorities a large, well-organised and motivated workforce to help them fulfil their statutory monitoring requirements. However, many of these participants have a genuine passion for environmental protection and citizen science, and may even possess species identification expertise and knowledge of local river issues and ecology rivalling that of experts (Waterton 2003; Pocock et al. 2014). Many of these volunteers are also keen to make additional useful contributions to river protection through further voluntary monitoring, restoration and

conservation work to fulfil local needs (Brooks et al. 2019). This may involve participating in the various citizen science schemes under the 'Riverfly Plus' umbrella, such as species surveys, outfall and misconnection surveys, algae monitoring and water chemistry monitoring. Other schemes volunteers take part in as part of 'Riverfly Plus' include the 'Extended Riverfly' and the 'Urban Riverfly', optional extensions to the ARMI protocol to increase the applicability of ARMI to incorporate the detection of impacts of fluctuations in water quality, low water flow and siltation, as well as urban-specific pollution, respectively, by analysing additional taxa alongside the standard eight (Fitch et al. 2018). All of these additional schemes help the Environment Agency meet their Water Framework Directive (WFD) objectives (Environment Agency 2019), and hence by encouraging and utilising local knowledge and interest in this way, ARMI is ensuring that river ecosystems around the country receive the highest levels of expertise, enthusiasm and protection available.

Likewise, through participation in ARMI, volunteers are able to learn how to use standardised monitoring protocols and equipment to obtain useful data, which they themselves can spatially and temporally compare with other data from the same river, region, or even across the country, alongside improving their knowledge of ecology and species identification (Brooks et al. 2019). Volunteers can also become more aware of environmental issues and their role in, and responsibility towards, the local environment (Storey et al. 2016; Fitch et al. 2018; Church et al. 2019), with regular nationwide communications and feedback, such as quarterly newsletters, articles in relevant publications and social media posts (http://www.riverflies.org/blogs/riverflies-news) providing opportunities for thought and discussion. Such democratisation has the potential to greatly advance citizen scientific and environmental literacy, as well as appreciation of the environment. This in turn can promote positive behaviour change towards the environment among citizens (Church et al. 2019; Thornhill et al. 2019) and build support amongst citizens for conservation and environmental activities (Latimore & Steen 2014). It also allows citizens to have positive significant engagement with, and influence on, governments and regulatory authorities regarding freshwater planning and legislation (Storey et al. 2016; Stepenuck & Genskow 2019).

Engagement with nature

ARMI also offers benefits for the wellbeing of individuals and communities. Specifically, it provides a means for individuals to immerse themselves into nature and aquatic environments (Dunkley 2019). Such engagement is reported to have various benefits for an individual's health and well-being (Pillemer et al. 2010), as well as their personal satisfaction and social welfare (Muirhead 2011). Furthermore, the local nature of the scheme can make ARMI a focus for community engagement and teamwork, which increases social capital (increases in trust, harmony and co-operation within a community) (Sultana & Abeyasekera 2008). As urbanisation across Europe increases over the next 30 years, access to urban freshwater may prove an essential strategy to address the negative health impacts of urbanisation and climate change (Dunkley 2019; Higgins et al. 2019) and therefore the importance of ARMI in facilitating engagement with nature is likely to only increase.

1.3 Challenges of ARMI

Like many citizen science schemes (Conrad & Hilchey 2011), ARMI faces difficulties in ensuring that the data it obtains is of high quality and quantity (Brooks et al. 2019). It is imperative that ARMI overcomes such difficulties to ensure there is enough reliable and accurate data so that the scheme is able to sufficiently and consistently fulfil its intended potential of complementing statutory routine monitoring.

ARMI has already introduced various strategies to avoid issues which would typically reduce the quantity of useable data collected through various means. For example, by bringing all concerned stakeholders under one standardised framework, issues such as different groups using different monitoring methods, or methods not suited to the goal of research (Di Fiore & Fitch 2016), are avoided. Similarly, by researching volunteer demographics and motivations (Isaacs 2017; Dunkley 2019), and ensuring that ARMI is appealing to citizens around the country and fulfilling their motivations (Roy et al. 2012), high levels of nationwide volunteer recruitment and retention are achieved (Brooks et al. 2019). In particular, volunteers remain engaged with the scheme due to the feedback and communication they are provided with by the regulatory authorities, as well as the sense of ownership they gain from the bottom-up nature of the project, and opportunities to connect with other volunteers on a local, regional and national level (Brooks et al. 2019). This enables high quantities of data to be collected, as well as spatial and temporal biases to be reduced. The involvement of regulatory authorities in approving the suitability of sites for volunteers to monitor (Brooks et al. 2019) also reduces spatial biases in the data.

Furthermore, through various mechanisms ARMI improves the reliability and accuracy of the data. For instance, by training volunteers to use a simplified version of a standardised and recognised method, and standardised equipment, ARMI avoids the use of biotic indices not well able to accurately or reliably assess the health status of the river (Di Fiore & Fitch 2016), as well as weakened methodological standards and lower-quality equipment (Cohn 2008). Furthermore, false reports of breaches (Royle 2004), which could result in regulatory authorities' time and resources being wasted, are avoided by having the regulatory authority set the trigger level to be sensitive to serious pollution. Such problems are also avoided through the score proofing system that is in place; when results are submitted to the river co-ordinator, he/she queries any unusual results, and before breaches are reported to the regulatory authority, a second sample is taken at the same site (Brooks et al. 2019).

Whilst the measures described above do go a long way to minimise data collection errors and improve the accuracy and reliability of the data, evaluating data is still an important step for regulatory authorities and researchers to have full confidence in their use of data gathered by the scheme (Connors et al. 2012; Bonney et al. 2014). Some evaluation of ARMI data has already been undertaken, with a recent analysis that compared nearly 5000 samples of professional BWMP and ARMI data showing that ARMI

data is comparable to professional monitoring (Brooks et al. 2019). Furthermore, initial pilot testing of the ARMI protocol suggested that with one day of training volunteers were able to reliably perform the standardised sampling, identify the eight taxa and analyse the data (Brooks et al. 2019).

However, while volunteers are advised at their training to strictly follow the prescribed ARMI protocol to ensure consistency and reliability of the data (STAR 2004; Edwards 2016), no analysis so far has considered variation in ARMI scores between different ARMI monitors, and how site water quality or inter-sampler differences (including differences in the sampling protocol used) could influence the scoring scheme (Brooks et al. 2019). Knowledge of this is essential so that, if necessary, the impacts of any inter-sampler and site water quality differences can be accounted for during use of the ARMI data (Clarke et al. 2002; Cooper et al. 2012), as well as to confirm how reliable the scheme is, and indicate whether and what improvements could be made to the protocol or training to increase reliability (Bonney et al. 2014; Ramos-Merchante & Prenda 2017; Tredick et al. 2017). Achieving a high level of reliability is critical so the data can be fully utilised by regulatory authorities. This includes for both the detection of one-off pollution incidents, and analysis of the long-term dataset for other trends, such as temporal and spatial trends in water quality, and changes in riverfly abundance and distribution (Brooks et al. 2019). Use of the data in this way has the potential to greatly improve river management (Edwards 2016) but also ensures the continuation of the scheme by attracting volunteers and funding, as the data is demonstrated to be impactful (Thornhill et al. 2019). Continuation of the scheme is important given the benefits of the scheme to regulatory authorities, volunteers and the environment previously described.

1.4 Aims and objectives

This study will therefore aim to determine the reliability of the ARMI scheme. It will achieve this by investigating the inter-rater reliability of ARMI scores generated by different volunteers for the same site, and the influence of inter-sampler and site water quality on score differences between different ARMI volunteers for the same site, as these factors have been previously found to influence riverine macroinvertebrate bioassessment results (Furse et al. 1981).

Specific inter-sampler differences that will be researched for their impact on score differences include whether or not volunteers follow the standard kicking and stone search method, differences in time taken to analyse the sample, and differences in size of the group they are working in (individuals or pair/multiples). These factors were chosen as they were found to vary greatly between volunteers, and due to the potential impact each of them may have on sample scores obtained. For example, subjectivity between volunteers regarding how to physically undertake a kick sample, or how to proportionally sample the different habitats, may result in certain taxa being over or under-collected in a sample, (Mackey et al. 1984; Kerans et al. 1992; Davies 2001; Haase et al. 2004a; Blocksom et al. 2008) as could not completing, or incorrectly completing, the stone search (Letovsky et al. 2012). Moreover, while

volunteers are advised to carry out their sampling in at least a pair, some volunteers do carry out sampling individually. Aside from safety concerns regarding lone fieldwork, sorting as an individual could result in reduced sorting effort, which may lead to specimens being missed and produce different scores compared to if there were two individuals sorting the sample. Similarly, while there is no set, recommended time for sample sorting, it has been previously observed that the longer spent sorting a macroinvertebrate sample, the more specimens that are likely to be found (Ettinger 1984; Vlek et al. 2006), which could also influence the scores obtained. However, after a certain amount of time there may be no, or only minimal, further gains to the score, and extra sorting is not necessary (Feeley et al. 2012). Hence, it will also be investigated whether differences in sample sorting times impact the score differences, and whether there is any advised minimum and maximum times for ARMI sample sorting.

Using the ARMI score achieved by the main investigator for each sample, the impact of site water quality on score differences between the main investigator and volunteer for the same site will also be investigated. Knowledge of this is essential as ARMI needs to be reliable regardless of whether it is used on poor (low ARMI scoring), moderate (medium ARMI scoring) or good (high ARMI scoring) water quality sites.

This study will focus on answering the following questions:

1. How reliable are i) overall and ii) taxa-specific ARMI scores obtained by different samplers at the same site?
2. Does site water quality influence differences in ARMI scores obtained by different samplers at the same site?
3. What methods do volunteers use to undertake their kick sample and stone search, and do specific inter-sampler differences (e.g differences in kick sample and stone search methodology, group size, sample sorting time) influence differences in i) overall and ii) taxa-specific ARMI scores obtained by different samplers at the same site?
4. Is there a minimum and maximum recommended time for sorting a sample?

The corresponding hypotheses are as follows:

1. Overall and taxa-specific ARMI scores obtained by different samplers at the same site are very reliable.
2. Site water quality does not influence differences in ARMI scores obtained by different samplers at the same site.
3. Volunteers follow the prescribed method for their kick sample and stone searches, and any specific inter-sampler differences (e.g differences in kick sample and stone search methodology, group size, sample sorting time) do not influence differences in i) overall and ii) taxa-specific ARMI scores obtained by different samplers at the same site.
4. There is a minimum and maximum recommended time for sorting a sample.

2. Methods

2.1 Study sites

Sampling was undertaken by the main investigator and 13 volunteer groups at one of the groups' usual monitoring sites located across Hillingdon, Buckinghamshire and Hertfordshire, UK (Figure 2; Table 2). Three sites ('Park Street Upstream' in St Albans; 'Dolittle Mill' in Redbournbury; 'Drop Lane' in Bricket Wood) are located on the River Ver, a 20km chalk stream running from its source at Kensworth Lynch to its confluence with the River Colne at Bricket Wood (Hertfordshire Life 2010). Two sites ('Springwell Lane Two' and 'Springwell Lane Upstream' in Rickmansworth) are situated the River Colne, a 18km chalk stream which flows from its source at North Mymms to its confluence with the River Thames at Staines-upon-Thames. A fast flowing, clear river, the sites analysed on this river are composed of Upper cretaceous chalk, overlain by sands, gravels and alluvium (CVRPP 2017), and run alongside Springwell and Stocker's Lakes, former gravel workings which are now sites of great ornithological significance (Baxter 2011). Another site ('Tykes Water' in Radlett) is located on Tykes' Water, a minor tributary of the River Colne running from Aldenham Reservoir to the River Colne at Colney Street (Radlett Neighbourhood Plan Steering Group 2019). One further site ('Ruislip') is found on the River Pinn, a 19 km stream originating in Pinner and flowing into the Frays River, a distributary of the River Colne, at Yiewesley (Mkandla 2018).

A further three sites ('Scotsbridge Mill' in Rickmansworth; 'Sarrat Mill Bridge' in Sarratt; 'Latimer Park' in Chesham) are situated on the River Chess, a 17.9km chalk stream running from its source at Chesham to its confluence with the River Colne at Rickmansworth (National Rivers Authority undated). A clear, fast flowing river, it flows through upper and middle chalk outcrops, overlain with gravel, and in some areas silt (National Rivers Authority Undated). The final three sites ('Doctor's Meadow' in Little Missenden; 'Denham Country Park' in Denham; 'Higher Denham' in Denham) are situated on the River Misbourne, a 27km chalk stream running from its source at Great Missenden to its confluence with the River Colne near Uxbridge (National Rivers Authority Undated). As a winterbourne, parts of the River Misbourne are prone to low flow (Bailey 2009), including both Doctor's Meadow and Higher Denham.

2.2 Data Collection

At each of the 13 sites, sampling was undertaken once a month for three months, at approximately the same time every month (with the exception of a few samples where the time gap was between 3-8 weeks due to availability/ weather). The overall sampling period lasted between late August- mid December 2018. For the majority of cases (n=28) the volunteer and main investigator collected and sorted at the same time, otherwise samples were taken within 24 hours of each other (n=11) so that

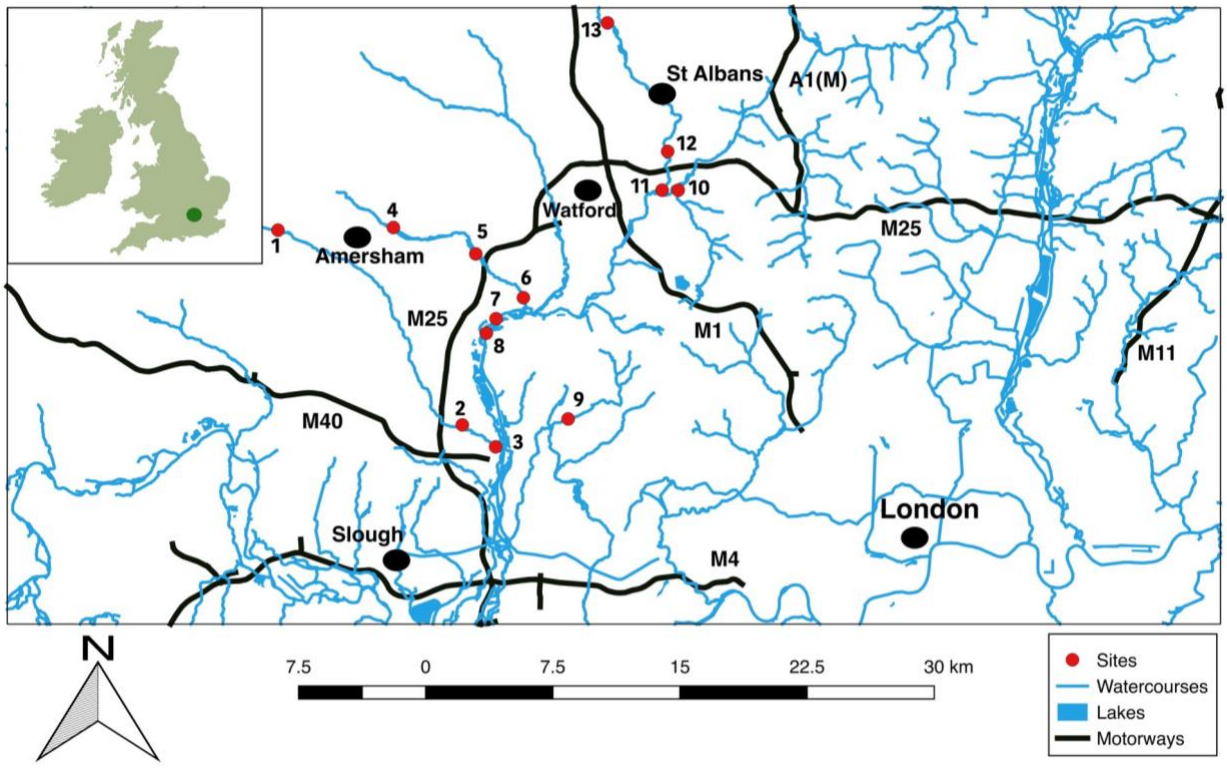


Figure 2. Location of sampling sites within Hertfordshire, Buckinghamshire and Hillingdon, UK. 1= Doctor's Meadow, 2= Higher Denham, 3= Denham Country Park, 4= Latimer Park, 5= Sarratt Mill Bridge, 6= Scotsbridge Mill, 7= Springwell Lane Upstream, 8=Springwell Lane Two, 9= Ruislip, 10= Tyke's Water, 11= Drop Lane, 12= Park Street Upstream, 13= Dolittle Mill.

Table 2. Site Names and grid references

Site Name	River	Grid reference
Park Street Upstream	Ver	TL 15005 03748
Dolittle Mill	Ver	TL 11454 11317
Drop Lane	Ver	TL 14678 01467
Springwell Lane Two	Colne	TQ 04302 93019
Springwell Lane Upstream	Colne	TQ 04877 93894
Tyke's Water	Tyke's Water	TL 15621 01471
Ruislip	Pinn	TQ 09146 87999
Scotsbrige Mill	Chess	TQ 06495 95133
Sarratt Mill Bridge	Chess	TQ 03687 97714
Latimer Park	Chess	SU 98828 99266
Doctor's Meadow	Misbourne	SU 9205699113
Denham Country Park	Misbourne	TQ 0485386346
Higher Denham	Misbourne	TQ 02887 87646

conditions remained relatively similar. In the case that sampling took place at the same time, approximately half of the samples were taken by the volunteer before the main investigator (n=13), and the other half after (n=15). This is to account for any impacts a freshly disturbed site may have on the score. Although collection and analysis took place at the same time, the volunteer and main investigator collected and sorted their sample independently.

Both the main investigator and all the volunteers used the standard recommended equipment advised in the FSC Riverfly Monitoring guide (specifically, standard kick sample net with 1.5m handle, 25cm frame and a 1mm net with 50cm depth, large collection bucket, large, white sorting tray, 8-section divider tray, turkey baster, magnifying glass/ hand lens). The main investigator also incorporated a stopwatch for timing the sample collection, and both the main investigator and a few volunteers also incorporated a spoon and fine paintbrush into their kit for picking out invertebrates during the analysis.

The main investigator followed the standard sampling procedure advised in the FSC Riverfly Monitoring guide. This consists of a three-minute kick sample, split proportionally between the different habitats available, kicking on the spot in numerous spots within a habitat, and sweeping through weed areas and vegetation, working across the river and progressively upstream. This is followed by a one-minute manual search of large liftable stones, which are wiped by hand in the water in front of the net. The sample is then washed, including removing large unwanted debris present in the sample e.g stones and leaves, and relevant taxa (cased caddis, caseless caddis, Ephemeridae, Ephemerellidae, Heptageniidae, Baetidae, Plecoptera and Gammarus) are picked out and sorted according to taxon in the segmented tray. The abundance of these invertebrates are then counted/estimated, and awarded a score, according to Table 1. The score for each taxon is added to give an overall score for the sample, with higher scores indicating sites have better water quality. If no (live) individuals of a certain taxon are found, that taxon is awarded a score of 0.

Both the main investigator and the volunteers have been trained in using this method at a 1-day workshop, and are expected to use it during sample collection and sorting. However, to account for inter-sampler differences, the main investigator also made notes on whether or not the volunteer followed the recommended method for completing their kick sample and stone search, and if not, details of the method they employed. Details of other inter-sampler differences, such as the number of volunteers in their group, and the time taken by both the main investigator and volunteer to sort their sample, were also recorded during the sampling event. The volunteers forwarded their scores and counts to the main investigator via e-mail after the sampling event.

2.3 Data Analysis

All analyses were carried out using SPSS version 25 (IBM Corp 2017). To analyse the inter-rater reliability of the scores, a single-rating, absolute agreement, one-way random effects ‘Intraclass

Correlation Coefficient' (ICC) model was calculated using the main investigator and volunteer scores of each of the 39 samples. This was carried out for both the overall scores, and taxa-specific scores. This reliability index was chosen as it is suitable for instances where different sets of volunteers assess different subgroups of sites (Koo & Li 2016). The ICC analysis was carried out in accordance with Aldridge (2015), and, as recommended by Koo and Li (2016), the ICC value and upper and lower bounds were reported; ICC values less than 0.5 indicate 'poor' reliability, values between 0.5 and 0.75 indicate 'moderate' reliability, values between 0.75 and 0.9 indicate 'good' reliability, and values over 0.9 indicate 'excellent' reliability.

To understand the impact of site water quality on score differences, i.e whether the difference between the volunteer and the main investigator's score varies as the main investigator's score does, a linear mixed model (LMM) was used. To determine the impact of inter-sampler differences on score differences, i.e whether the score difference between the volunteers and main investigator varied depending on: whether or not the volunteers followed the same method as the main investigator for the i) kick sample and ii) stone search, iii) whether or not they completed their monitoring as an individual or a pair/group, and iv) difference in time taken to analyse the sample, a separate LMM was created. An LMM was chosen for these analyses to account for the correlation within the data due to the repeated measures nature of the sampling (Fitzmaurice & Laird 2015) (i.e the same volunteers were used three times over), and because it accounts for missing data (Fitzmaurice & Laird 2015) (i.e if the main investigator was not able to record information on inter-sampler differences) (Maxwell et al. 2017).

To determine whether there was a certain time point during analysis at which further score gains became unlikely, a Z-score was calculated for each overall sample score using the mean and standard deviations of the three samples of each site, and compared graphically with the time taken to analyse that sample.

3. Results

3.1 Reliability of the ARMI protocol

On average volunteers achieved a score of 9.7, with a range of 2-16, while the main Investigator achieved an average score of 9.3 with a range of 3-16. There was 'good' significant intra-class correlation between the volunteers' and main investigator's overall scores (Figure 3, Table 3), with sites of a variety of water qualities used in this analysis (ranging from 3-16; poor-excellent water quality), but the difference in scores between the main investigator and the volunteers not statistically significantly varying across the main investigator's range of scores (Figure 4, $F_{(12,17.973)}= 0.866$, $p>0.05$).

Furthermore, the intra-class correlation between the main investigator's and the volunteers' taxa-specific scores was significant ($p<0.05$; Table 3). The ICC category was 'excellent' for mayfly (Table 3; Figure 5), 'good' for Gammaridae and Heptageniidae (Table 3; Figure 6), 'moderate' for cased caddisfly and Baetidae (Table 3; Figure 7), and 'poor' for Ephemerellidae and caseless caddisfly (Table 3; Figure 8). No Plecoptera were found in any samples by either the main investigator or the volunteers, so no analyses could be undertaken on this taxon.

Table 3. ICC values for the overall scores and each taxon individually, along with lower and upper ICC value bounds, and the p-value, summarised with the overall ICC category.

Taxon	ICC value	Lower Bound	Upper Bound	P-value	Overall ICC category
Overall	0.832	0.704	0.908	<0.05	Good
Mayfly	0.936	0.882	0.966	<0.05	Excellent
Gammaridae	0.871	0.768	0.930	<0.05	Good
Heptageniidae	0.802	0.655	0.891	<0.05	Good
Cased caddisfly	0.505	0.232	0.705	<0.05	Moderate
Baetidae	0.618	0.382	0.779	<0.05	Moderate
Ephemerellidae	0.289	-0.023	0.550	<0.05	Poor
Caseless caddisfly	0.474	0.193	0.684	<0.05	Poor

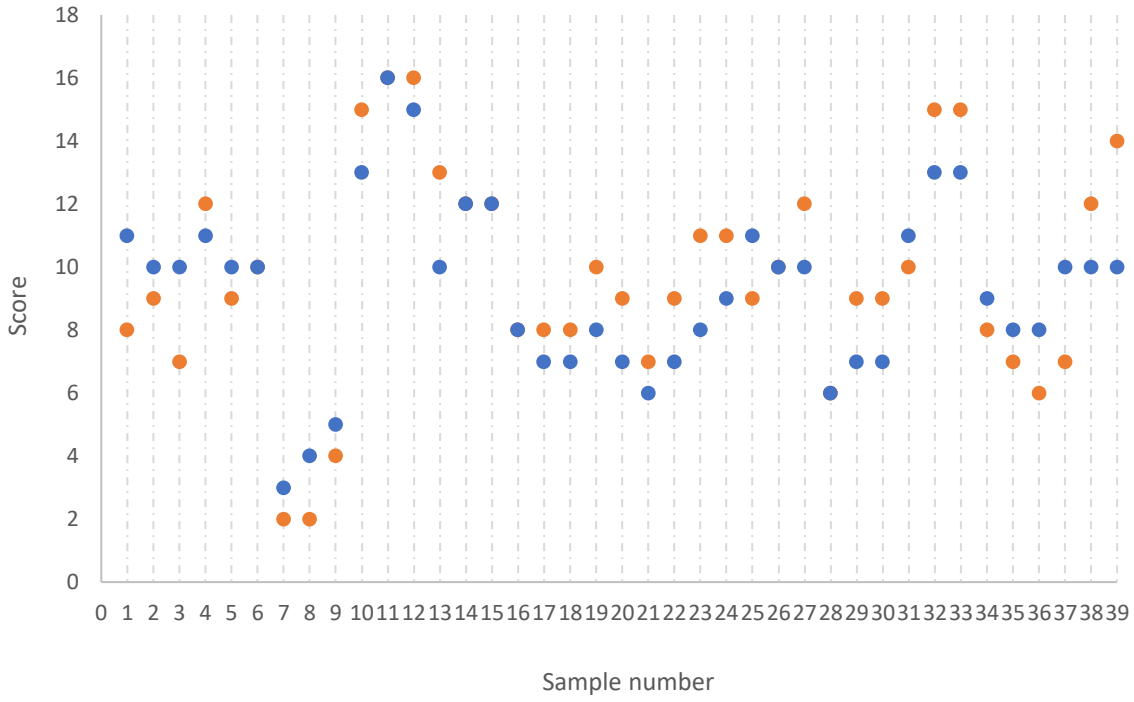


Figure 3. ICC plot showing the overall scores obtained by both the main investigator (blue) and the volunteer (orange) for the 39 pairs of samples. If only one dot is shown per sample it indicates the scores are identical.

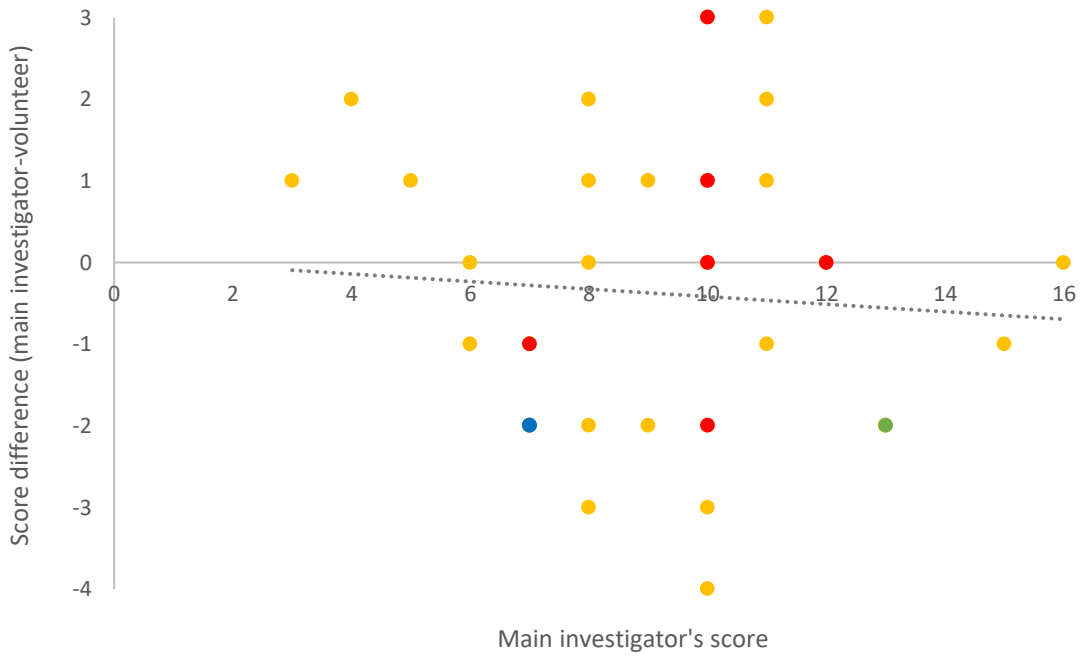


Figure 4. Relationship between the main investigator's score and the difference between that score and the volunteers' for each of the 39 pairs of samples. Yellow markers indicate that that score difference occurred for that main investigator score on one occasion, the red marker indicates it occurred on two occasions, the green marker indicates it occurred on three occasions, and the blue marker indicates it occurred on four occasions. Dotted regression line shown (R^2= 0.005, Y= 0.04-0.05x).

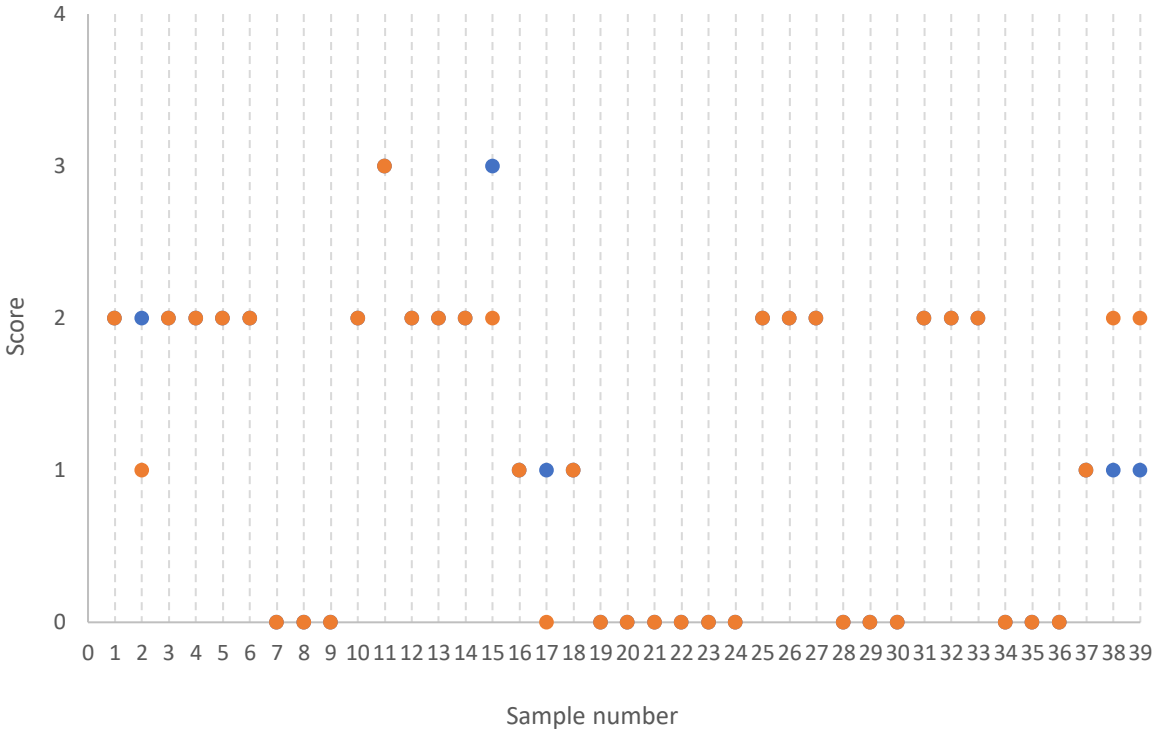


Figure 5. ICC plot showing the scores obtained for mayflies by both the main investigator (blue) and the volunteer (orange), for each of the 39 pairs of samples. If only one dot is shown per sample it indicates the scores are identical.

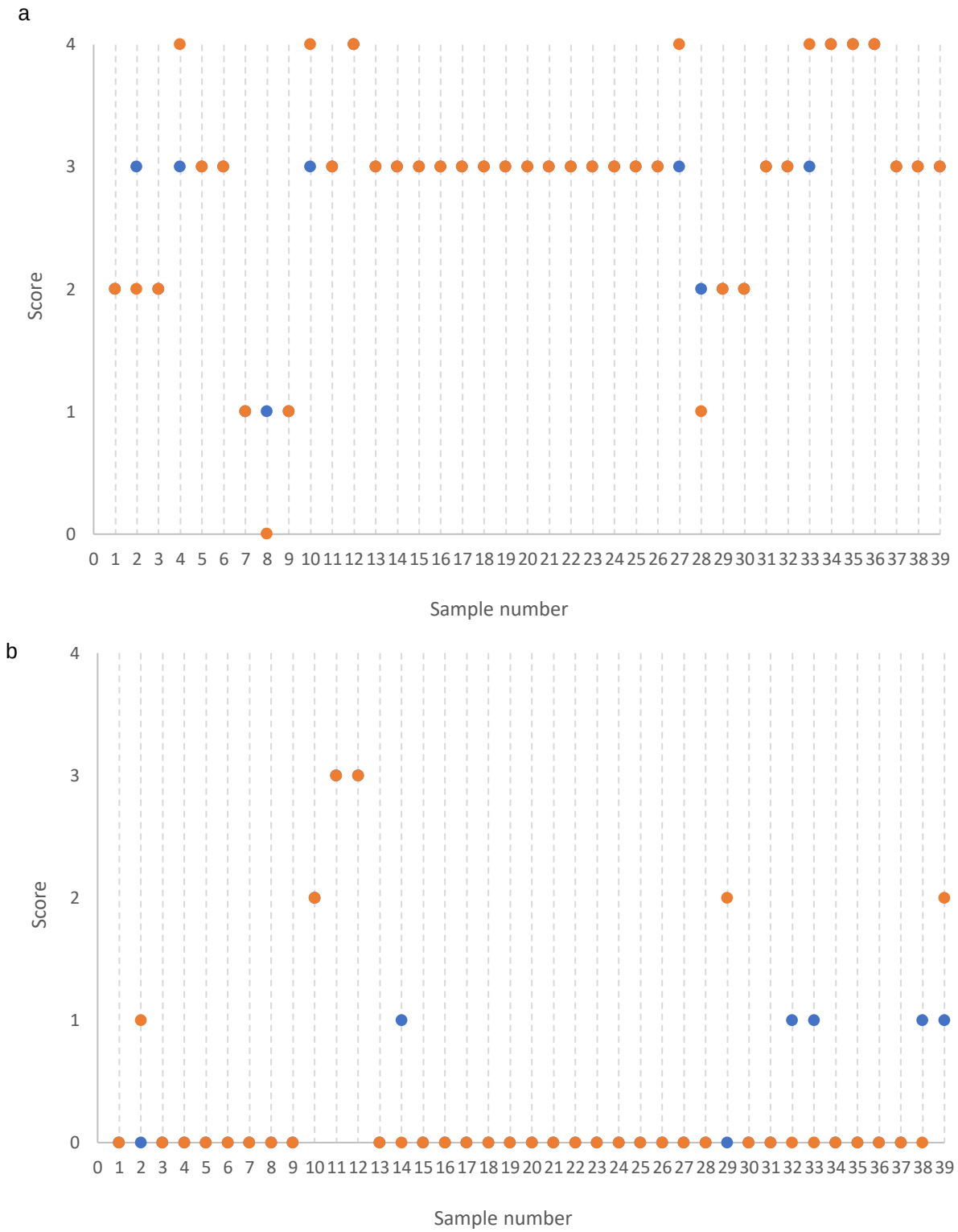


Figure 6. ICC plot showing the scores obtained for a) Heptageniidae and b) Gammaridae by both the main investigator (blue) and the volunteer (orange), for each of the 39 pairs of samples. If only one dot is shown per sample it indicates the scores are identical.

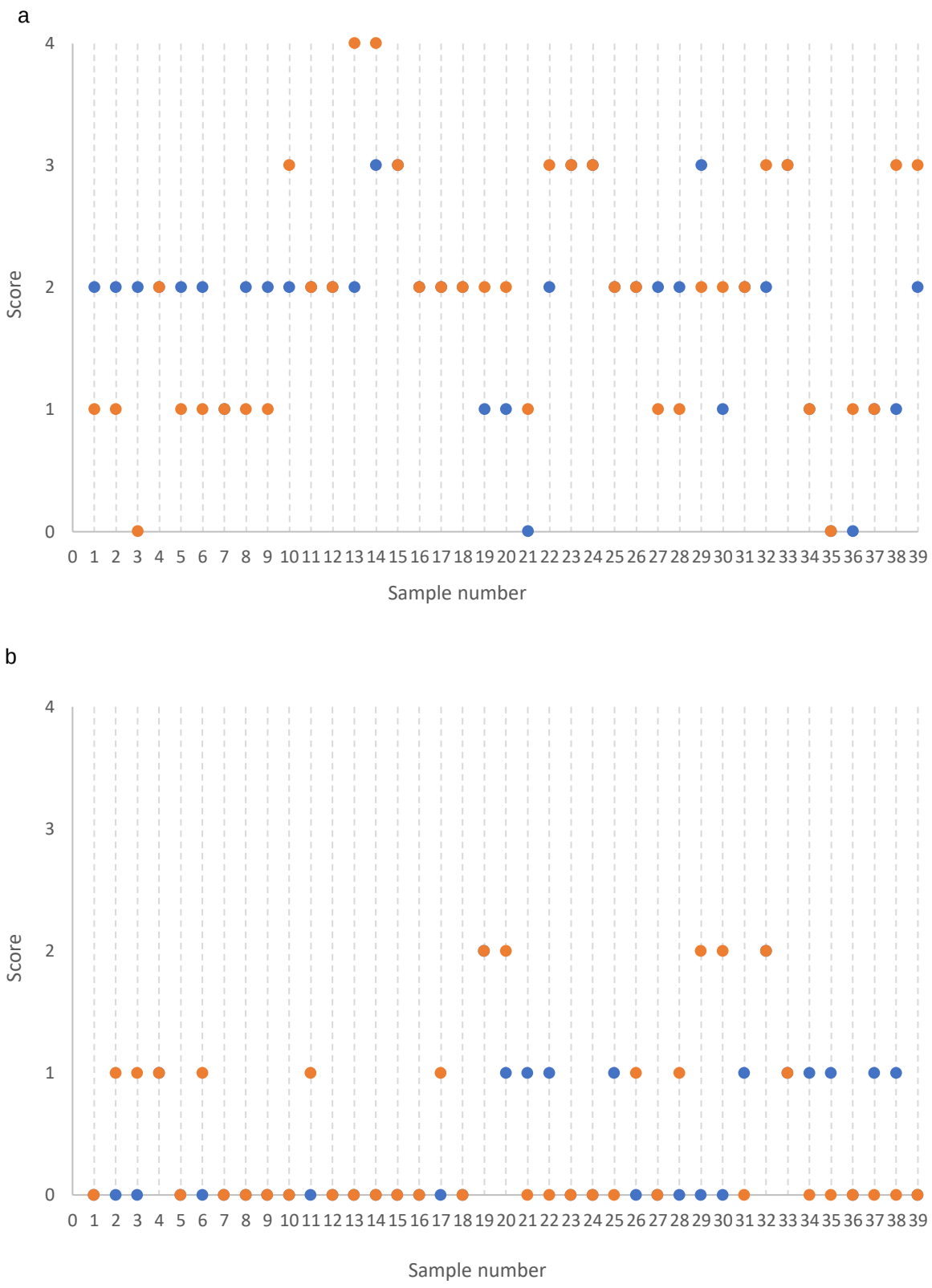
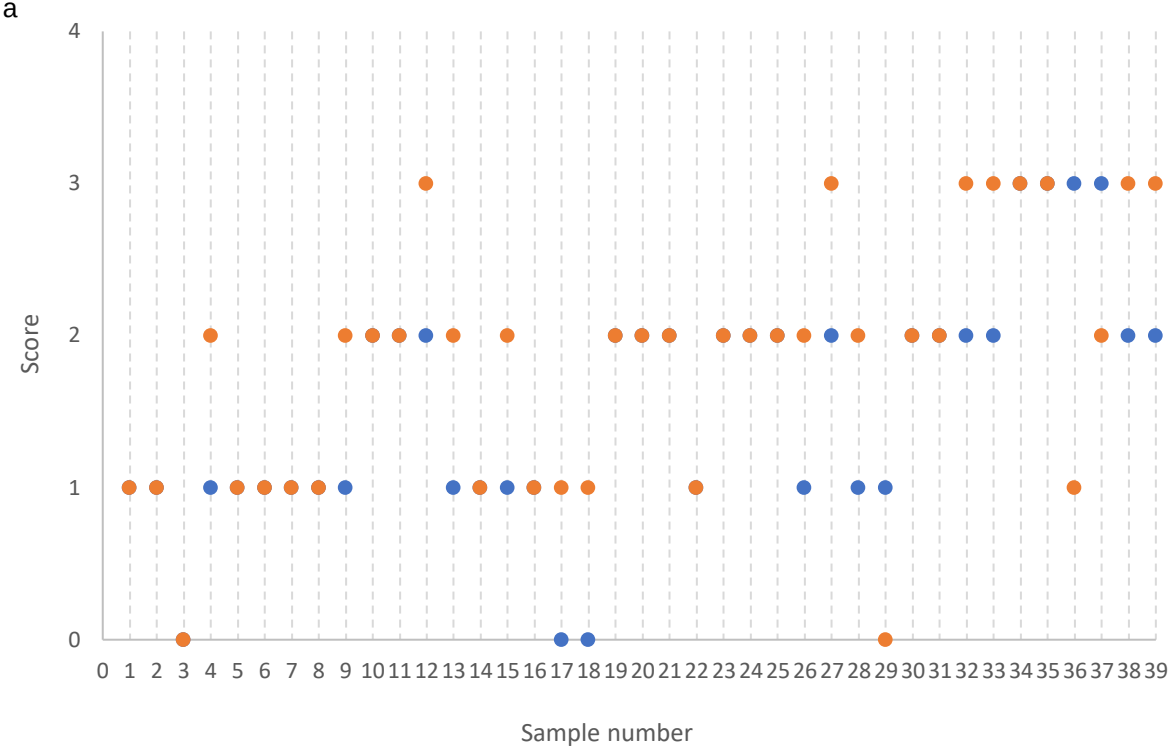


Figure 7. ICC plot showing the scores obtained for a) cased caddisfly and b) Baetidae by both the main investigator (blue) and the volunteer (orange), for each of the 39 pairs of samples. If only one dot is shown per sample it indicates the scores are identical.

a



b

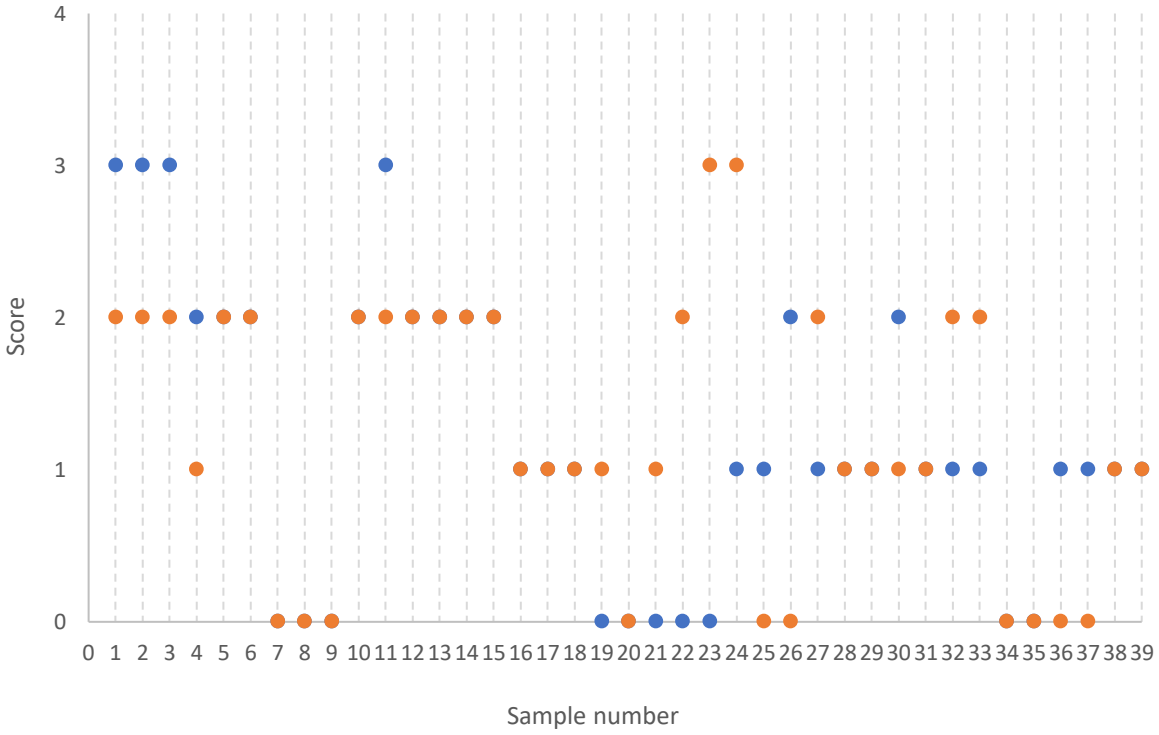


Figure 8. ICC plot showing the scores obtained for a) Ephemerelliidae and b) caseless caddisfly by both the main investigator (blue) and the volunteer (orange), for each of the 39 pairs of samples. If only one dot is shown per sample it indicates the scores are identical.

3.2 Impact of inter-sampler differences on the scores

Of the samples the volunteers competed, 22 were undertaken as a pair/ group, while 17 were undertaken by individuals. 20 samples were undertaken using the standard stone search method (i.e washing stones in front of the net for one minute), with 11 being undertaken using an alternative method (specifically, counting target invertebrates directly on a stone, or not completing a stone search when stones were available to do so). 18 samples were undertaken using the standard kicking method (i.e kicking in many spots throughout the river and proportionally in different habitats, working progressively upstream for three minutes). 10 samples used a variation of the standard kicking method (i.e kicking and walking up the river at the same time, or only kicking in only two or three different spots in the river overall, including those who did not proportionally sample all habitats, with certain habitats such as vegetation and silt avoided in some cases). Data on difference in time taken by the main investigator and volunteer to sort their sample was gathered for 28 of the samples (Figure 9).

According to the LMM analysis, none of these inter-sampler differences were found to influence overall score differences, nor score differences for Baetidae, caseless caddisfly, Ephemeridae, Gammaridae nor Heptageniidae. For cased caddisfly, however, it was found that samples undertaken with a non-standard kick sample method had a significantly higher score difference with the main investigator than samples undertaken with the standard kick sample method ($F_{(1,6.8)}$=12.5, $p<0.05$). Similarly, samples undertaken by a group for both cased caddisfly ($F_{(1,14.6)}$=6.83, $p<0.05$) and Ephemerellidae ($F_{(1,14.7)}$=6.7, $p<0.05$) had a significantly larger score difference with the main investigator than samples undertaken by individuals. Moreover, for Ephemerellidae, samples with a larger sorting time difference than the main investigator had more similar scores to the main investigator compared to samples with less of a time difference ($F_{(1,22.4)}$=4.7, $p<0.05$; Figure 10), with the graph suggesting that the main investigator needed longer to sort their sample. There was, however, no time point during sorting of the collected sample at which further score gains became more or less likely (Figure 11).

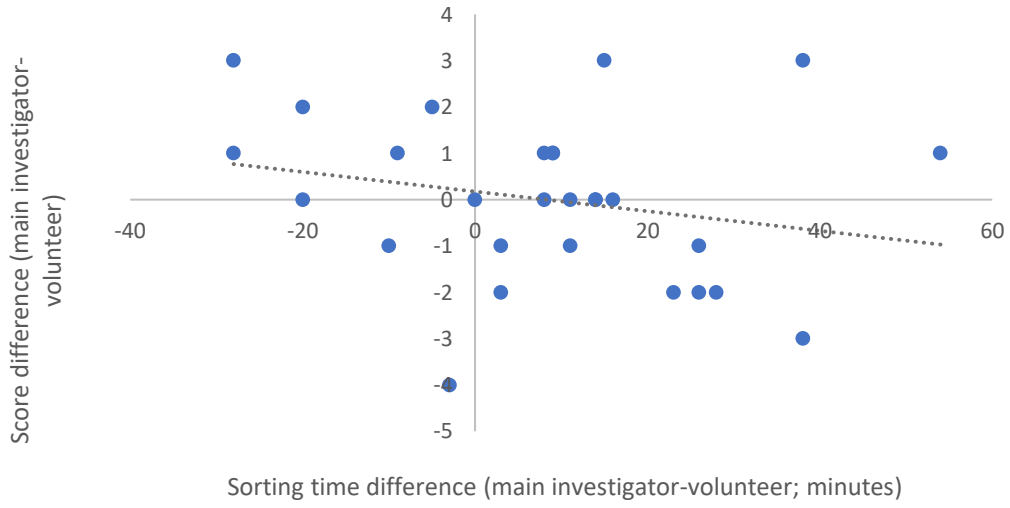


Figure 9. Relationship between the difference in sample sorting time between the main investigator and volunteer, and the score difference the main investigator and volunteer for overall sample scores, for 26 of the 39 samples. Dotted linear trendline shown (R^2= 0.055, Y= -0.021x + 0.175).

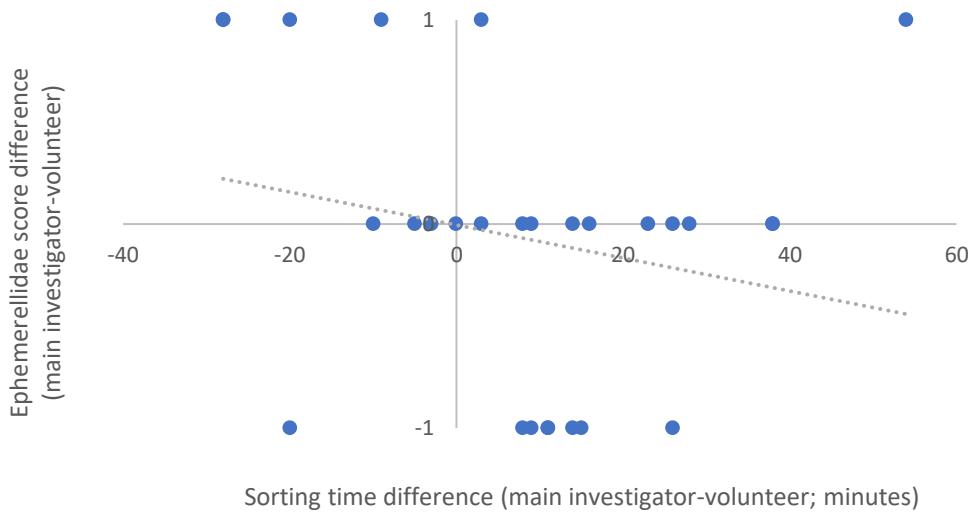


Figure 10. Relationship between the difference in sample sorting time between the main investigator and volunteer, and the score difference between the main investigator and volunteer for Ephemerellidae, for 26 of the 39 samples. Dotted linear trendline shown (R^2= 0.05, Y= -0.008x + 0.005).

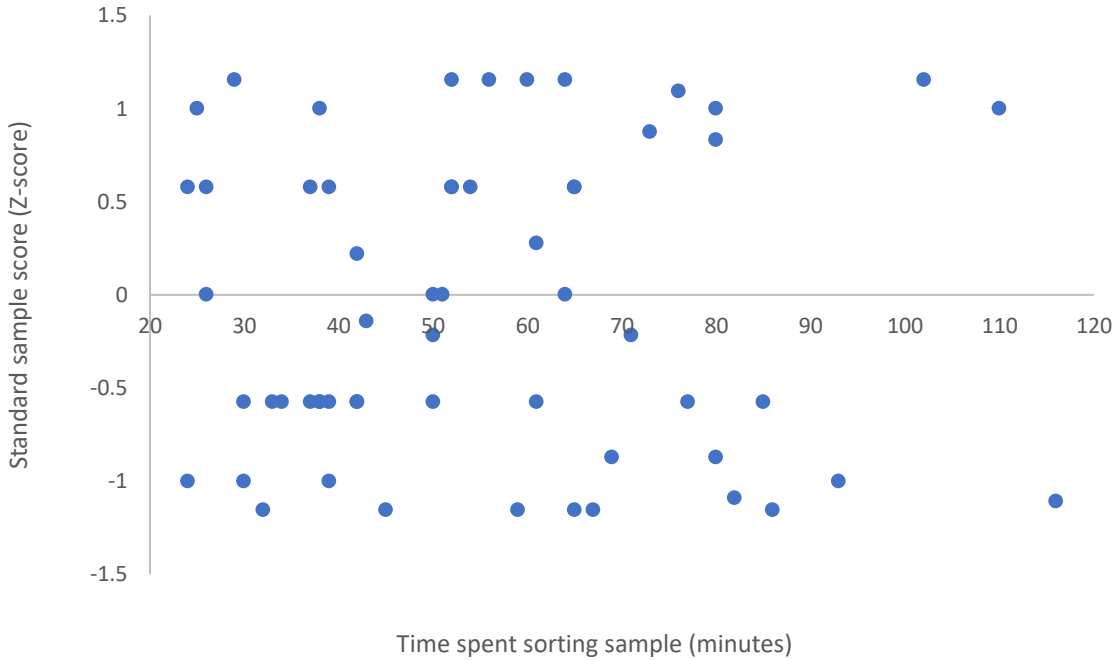


Figure 11. Relationship between the time taken to sort the sample in minutes, and the standard score (Z-score) obtained for that sample, for 58 of the 78 samples undertaken overall by both the volunteers and the main investigator.

4. Discussion

4.1 Overall reliability

ARMI was found to have 'good' inter-rater reliability overall, and while there were differences in sizes of the groups, techniques that volunteers used to complete the kick sample and stone search, and time taken to sort the invertebrate samples, none of these inter-sampler differences were found to significantly influence the overall score differences between volunteers. This agrees with previous research which showed that BMWP sampling values, on which the ARMI is based, are not impacted by inter-sampler influences (Clarke et al. 2002). Much like other ecological and water quality citizen science schemes which have been carefully designed with the oversight of professionals and with strategies put in place to mitigate data quality issues (Buckland-Nicks et al. 2016; Thornhill et al. 2018; Brown & Williams 2019), these results concur that the use of simplified but standardised methods and equipment (Edwards 2016; Rose et al. 2016; Storey et al. 2016; Franca et al. 2019; Seelen et al. 2019; Weigelhofer et al. 2019), a simplistic identification and scoring system (Rose et al. 2016; Storey et al. 2016; Franca et al. 2019), and a brief but comprehensive training programme (Edwards 2016; Ratnieks et al. 2016; Storey et al. 2016; Thornhill et al. 2018; Franca et al. 2019; Oti et al. 2019; Seelen et al. 2019) is a robust, user-prone, reliable way to produce data. The sites used in this study also had a wide range of water quality levels (scores ranged from 3-16; poor-excellent water quality), but the finding that the score difference between the main investigator and volunteers are not influenced by the main investigator's score indicates ARMI is reliable for sites of all water quality levels. Taking these outcomes combined with recent findings that ARMI scores correlate well with professional monitoring data (Brooks et al. 2019), it is therefore fair to say that ARMI is fulfilling its intended role of reliably complementing routine statutory monitoring.

That ARMI reliably complements routine statutory monitoring has various implications for the scheme and data it produces. This includes that reports of trigger level breaches should be taken seriously and acted on by regulatory authorities, and that datasets gathered so far would be of use for investigating long-term trends, such as to understand temporal and spatial trends in water quality, and changes in riverfly abundance and distribution (Brooks et al. 2019) of taxa which had high levels of reliability, without the need to account for site water quality or inter-observer differences during analysis (Cooper et al. 2012). These findings also indicate that we should have confidence in the development and use of the Riverfly Plus schemes based on ARMI, such as the 'Extended Riverfly', 'Urban Riverfly', as well as the 'Restoration Assessment Initiative', a suggested adaptation to the ARMI to monitor the impacts of river restoration (Huddart et al. 2016).

Moreover, given the benefit that such reliable data provides for regulatory authorities in supporting their statutory monitoring, it is recommended for regulatory authorities to continue supporting the delivery and improvement of the scheme. This should include following suggestions made by Fitch et al. (2018), such as providing resources to EA ecology contacts so they can provide consistent and dedicated

support to ARMI and its volunteers (such as assisting with training programmes and attending local information meetings), providing further training for RP accredited trainers, and raising the profile of ARMI in Catchment Partnerships. Following further propositions made by Fitch et al. (2018), such as for the EA to always provide timely feedback to ARMI contacts on the outcomes of trigger level breach investigations, as well as continuing and increasing funding for ARMI, given that funding is a key limitation of the longevity of many citizen science programmes (Owen & Parker 2018), are also advised.

The finding that ARMI data is so reliable also suggests it is worth persisting with efforts to expand the volunteer base to enhance data collection across the UK, particularly in areas where ARMI activity is low, such as Kent, Sussex, Norfolk and West Midlands (Fitch et al. 2018). The potential for expansion is made substantially easier by the finding that ARMI is appropriate for sites of poor, moderate and high water quality. This is because this factor makes the scheme accessible for volunteers regardless of the water quality of their local river or site, which will also help to further reduce spatial bias in the data, a common issue with citizen science programmes (Thornhill et al. 2019). After modification to include locally relevant indicator taxa (Blijswijk et al. 2004) ARMI may even offer a viable option as a biotic water quality citizen science scheme in countries where one is not currently in existence. Not only would use of a well-structured and an already tried and tested water quality citizen science programme such as ARMI increase the likelihood of the data being of high quality and hence being used by regulatory authorities (Gouveia et al. 2004; Conrad & Hilchey 2011; Tredick et al. 2017), using common methods and platforms increases project cost-effectiveness and study scalability (Thornhill et al. 2018).

4.2 Taxa-specific reliability

Although the overall scores achieved a high inter-rater reliability and were not influenced by inter-sampler differences, this was not always the case taxa-specifically. For instance, cased caddisfly and Baetdiae achieved ‘moderate’ inter-rater reliability, while Ephemerellidae and caseless caddisfly achieved only ‘poor’ inter-rater reliability, with score differences between the volunteer and main investigator for cased caddisfly found to be influenced by differences in kick sample method and group size, and Ephemerellidae by differences in group size and time differences. Therefore, improvements to the protocol and training are necessary to rectify these issues, so that the overall ARMI reliability levels can be improved, and that ARMI data can be used to analyse changes in the distribution and abundance of these taxa.

i) Kick sampling technique

For cased caddisfly, samples undertaken without using the standard kicking protocol had significantly larger score differences with the main investigator compared to samples for which the standard method was used. Although this analysis is not able to reveal whether using the non-standard kick sample method caused scores to be higher or lower than the main investigator and other volunteers using the

standard method, as the Z-scores could not be calculated for many of the 39 cased caddisfly-specific scores, the alternative methods used by the volunteers in this study can impact the collection of cased caddisfly in a variety of ways. Kicking and walking upstream at the same time for example, as used by some volunteers, may cause less cased caddisfly to be collected, as the net is unlikely to stay on the riverbed at all times, and disturbed specimens may be swept under the net as opposed to into it. Similarly, most individual invertebrates are dislodged from the substrate in the first few seconds of kicking (Mackey et al. 1984), and therefore volunteers employing the alternative technique of kicking in a few spots for a minute each may not be collecting as many cased caddisfly as volunteers who sample more spots for less time. Moreover, samples undertaken without proportional sampling of the available habitats may have biased the collection of cased caddisflies, as has also been found elsewhere (Blocksom et al. 2008). This is because benthic macroinvertebrates, including specifically Trichoptera (Wallace 1991), are not evenly distributed across the different stream habitats, and rather occur in higher and lower abundances in particular habitats and patches (Carter & Resh 2001). Over or under-sampling of these habitats and patches may therefore cause cased caddisfly to be over or under-represented in the sample.

That some volunteers are not using the standard kick sample method implies not all volunteers are being taught the correct method, or that there is a disconnect between what volunteers are being taught at their training, and what they are taking away from it. As using the incorrect kicking method impacts cased caddisfly inter-rater reliability, it is vital that only the standard method is used. Given that volunteers can be motivated to follow protocols when there is the prospect that their data will be valuable and used by regulatory authorities (Pocock et al. 2014), it is not inevitable that volunteers should use non-standard methods. Rather, clearly demonstrating the correct kick sample protocol to volunteers at their training and emphasising the importance of following this standard method to improve the reliability and accuracy of the data, so that the data can be fully utilised by regulatory authorities to make environmental improvements, will likely motivate volunteers to ensure they fully understand and carry out the standardised method. Specifically, as some volunteers were noticed to actively avoid silty habitats and vegetation, the FSC Riverfly Monitoring guide and training could be improved to explicitly teach volunteers how to correctly sample from these habitats without clogging up their nets, so that volunteers are more likely to proportionally sample all habitats present in the river. This should be based on the procedures outlined in sections 7.3b and 7.3d, respectively, by STAR (2004). Furthermore, as the sampling practice during the training day is undertaken in small groups, not all volunteers may have an opportunity to practice their kicking technique. It is also noticeable that, post-training, there is no further formal opportunities to give feedback to volunteers on their sampling technique. Consequently, it may also be worth considering introducing a new standard for the local river co-ordinator to attend the first sample with each new volunteer at their site, so that they can review and advise on the sampling methods used by the volunteer(s). Regulatory authorities are encouraged to make funding and resources available for this provision.

Additionally, even though the stone search technique used was not observed to significantly impact overall or taxa-specific score differences for the same site, some volunteers did fail to carry it out. Therefore, it should also be emphasised to volunteers, both in the training and the FSC Riverfly Monitoring Guide, that even when they suspect their stone search will not yield anything it should still be undertaken to maintain consistency (STAR 2004).

ii) Group size

Differences in cased caddisfly and Ephemerellidae scores are also influenced by differences in group size, with samples collected and sorted by groups having larger score differences with the main investigator compared to samples collected and sorted by individuals. Although this study cannot reveal whether groups achieved higher or lower scores than individuals, because the Z-scores could not be calculated for many of the 39 Ephemerellidae and cased caddisfly scores, it may be that due to their lower sample sorting effort, individuals are not noticing all specimens, although future research is needed to verify this. In the meantime, given that differences in group size (specifically between individual volunteers and multiple volunteers) do influence score differences, it is worth considering encouraging volunteer units to be of a similar size. As the current standard is to monitor in at least a pair, it is therefore recommended to advise individuals, unless really unavoidable, to also carry out their research in at least a pair. This will hopefully make Ephemerellidae and cased caddisfly results undertaken by different volunteers more comparable to one another.

iii) Other factors

Baetidae and caseless caddisfly also achieved low inter-rater reliability scores, however none of the factors examined in this study were found to significantly influence score differences for these taxa. It is therefore likely that other factors not examined here are responsible for this. This may include the observational finding that some volunteers did not follow the prescribed protocol of washing their sample and removing debris, which could impact scores as the more vegetative matter and debris in a sub-sample, the more difficult identification and estimating invertebrate abundance can be, particularly for non-experts (STAR 2004). However, not enough samples were undertaken without sample washing to allow for a full analysis based on the data collected here, and therefore future research will need to investigate whether this factor is also a common issue in the wider volunteer population, and whether it influences score differences and inter-rater reliability.

An alternative explanation for the low inter-rater reliability of these taxa may be differences in the abilities of volunteers to identify them, with some volunteers expressing concern about their own identification skills, and previous research showing identification skills can significantly influence the results of macroinvertebrate assessments (Furse et al. 1981; Haase et al. 2004b). Difficulties in identification may be due to the number of different species that exist for Baetidae and caseless caddisfly (Wallace 2003; Macadam 2016), as well as the wide size ranges of these taxa (Bouchard

2004a, b, c), but lack of explicit size guidance in the guide and training presentation. Differences in volunteers' identification skills are likely to be caused by differences in the quality of training they received, as well as differences in their level of previous experience in macroinvertebrate identification and sample sorting (Haase et al. 2004b; Cooper et al. 2012) (both in and outside of the ARMI scheme). Therefore, future work may look to investigate how well volunteers can identify the target taxa, and to what degree differences in identification skills impact the inter-rater reliability and score differences. This could be achieved by having different volunteers sort the same sample independently, and comparing their results.

If found to be necessary, these analyses will allow appropriate amendments to the training and protocol to be put in place. Such amendments may range from ensuring the importance of sample washing is emphasised to volunteers, improving the identification guidance and quality of sites used to train volunteers, or even introducing a compulsory post-training test to evaluate their identification skills (Louw et al. 2018).

This study was also unable to measure the inter-rater reliability, and the impact of inter-sampler differences on score differences, for Plecoptera, as none were found in any of the samples. Therefore, future research should look to analyse this by comparing results gathered by volunteers at sites where Plecoptera are known to be present.

4.3 Maximum and minimum recommended sample sorting times

It was found that there was no time point during sample sorting at which score gains became more or less likely. This is perhaps because the time required to sort a sample also varies depending on the volume of the sample load collected (Friberg et al. 2006), which itself can vary with level of vegetation and debris collected (Feeley et al. 2012), as well as stream substrate (Haase et al. 2004b). Similarly, sampling and identification experience, pleasant weather conditions and high levels of illumination can make spotting and differentiating taxa easier (Haase et al. 2004b), which in turn may speed up sorting. Taking this into consideration, and given the high level of overall inter-rater reliability achieved in this study without sample sorting time limits or recommendations, it is therefore recommended that there is no need to introduce minimum and maximum recommended sample sorting times. Rather, volunteers should continue to spend as long as they feel appropriate sorting to get an adequate overview of their sample. This is supported by the fact that for Ephemerellidae, score differences are significantly reduced for volunteers' samples undertaken with a larger sample sorting time difference compared to the main investigator, as in this case differences in sample sorting time actually helped reduce score differences.

5. Conclusion

This study has shown that the overall scores produced by the ARMI scheme are reliable and not influenced by site water quality or inter-sampler differences. This has various implications for the scheme and its data, including that reports of pollution should be taken seriously and acted on by regulatory authorities, and that datasets gathered thus-far should be analysed to understand temporal and spatial trends in water quality. Regulatory authorities are also advised to continue supporting, funding and helping with the expansion of the scheme. It is recommended, however, to increase the inter-rater reliability of Ephemerllidae and cased caddisfly, which are 'poor' and 'moderate' respectively, that the standard kick sampling technique, and the importance of following it, is better clarified to volunteers in the FSC Riverfly Monitoring Guide and training. It is also suggested that volunteers are provided with further opportunity for feedback on their sampling technique, and that unless avoidable, volunteer units are composed of at least sampling pairs as opposed to individuals. Further work is also required to understand and rectify the causes of the low inter-rater reliability of Baetidae and caseless caddisfly, which may even help increase the overall inter-rater reliability from 'good' to 'excellent'. There is, however, no recommended minimum and maximum sample sorting times, so volunteers spend as long as they feel appropriate sorting to get a good overview of their sample.

6. Auto-critique

My initial motivation for this topic was sparked by my interest in citizen science and environmental volunteering, and the opportunity to undertake a project that had the potential to demonstrate the worth of citizen science programmes really appealed to me. The high uptake and success of the ARMI scheme, but relatively little research into the quality of data produced from it, and the motivation and enthusiasm of the volunteers I reached out to to participate, made ARMI the perfect basis for my research. Overall, this study has been useful to highlight that the data gathered by ARMI is reliable and that while volunteers don't always follow the protocol exactly, the scoring scheme and methodology is robust enough that these differences in methodology do not influence the scores. Hopefully, this can be used to impress upon regulatory authorities that they should continue to use the data and support the scheme, and motivate organisers to continue with the expansion of the volunteer base. It has also highlighted areas where strategies are needed to improve reliability, and suggested steps that could be implemented to address this.

However, this study focuses mainly on how difficulties at the sample collection stage may influence the scoring system, and does not explicitly consider how differences in volunteers' identification abilities may influence the scores, and the causes for differences in volunteers' identification abilities. In hindsight it would have been useful to also research this, for instance by having different volunteers sort the same sample, and determining whether differences in identification abilities impact score differences more than differences at the sample collection stages. While data was collected on the experience of ARMI volunteers in the scheme, it was not possible to quantify or control for identification experience outside the scheme, e.g anglers, wildlife enthusiasts. Therefore, future research could improve on this by finding a way to quantify and control volunteers' entire previous experience in macroinvertebrate identification so the impact of such experience on identification abilities and score differences can be assessed. This would allow resources to be allocated to improving the area that would have the most impact on improving reliability. This study also only considers chalk streams; considering other types of streams may also be useful to verify the reliability of the protocol for streams found throughout the UK.

7. References

Abassi T, Abassi S 2012. Water Quality Indices. Elsevier. 384pp.

Aldridge V. 2015. Reliability Assessment Using SPSS SPSS Users Conference, University of York.

Bailey H. 2009. River Misbourne- Notes on the Geology. http://www.misbourneriveraction.org/node/20:7pp. (accessed 7/7/2019.

Bartle W. 2018. Chalk Stream volunteers help monitor Lincolnshire's rare waterways, https://www.lincolnshire.gov.uk/news/chalk-stream-volunteers-help-monitor-lincolnshires-rare-waterways/132971.article. (accessed 7/7/2019.

Baxter A. 2011. Colne River Valley. London's Natural Signatures: The London Landscape Framework / January 2011:32-35.

Blijswijk W, Coimbra C, Graca M. 2004. The use of biological methods based on macroinvertebrates to an Iberian stream (Central Portugal) receiving a paper mill effluent Limnetica **23**:307-314.

Blocksom KA, Autrey BC, Passmore M, Reynolds L. 2008. A comparison of single and multiple habitat protocols for collecting macroinvertebrates in wadeable streams. Journal of the American Water Resources Association **44**:577-593.

Bonney R, Shirk JL, Phillips TB, Wiggins A, Ballard HL, Miller-Rushing AJ, Parrish JK. 2014. Next Steps for Citizen Science. Science **343**:1436-1437.

Boren Z, Scott R. 2018. Inspections and pollution tests drop as Environment Agency sheds thousands of staff. Unearthed, https://unearthed.greenpeace.org/2018/12/08/environment-agency-pollution-inspections-cuts-rivers/. (accessed 7/72019

Bouchard R. 2004a. Aquatic Invertebrates. Pages 9-33. Guide to aquatic invertebrates of the Upper Midwest. Water Resources Center, University of Minnesota, St Paul, MN.

Bouchard R. 2004b. Ephemeroptera (mayflies). Pages 47-62. Guide to aquatic macroinvertebrates of the Upper Midwest. Water Resources Center, University of Minnesota, St Paul, MN.

Bouchard R. 2004c. Trichoptera (Caddisflies). Pages 115-135. Guide to aquatic macroinvertebrates of the Upper Midwest. Water Resources Center, University of Minnesota, St Paul, MN.

Brooks S, Fitch B, Davy-Bowker J, Codesal S. 2019. Anglers' Riverfly Monitoring Initiative (ARMI): A UK-wide citizen science project for water quality assessment Freshwater Science **38**:270-280.

Brown ED, Williams BK. 2019. The potential for citizen science to produce reliable and useful information in ecology. Conservation Biology **33**:561-569.

Buckland-Nicks A, Castleden H, Conrad C. 2016. Aligning community-based water monitoring program designs with goals for enhanced environmental management. Jcom-Journal of Science Communication **15**.

Cairns J, Pratt J. 1993. A history of monitoring using benthic macroinvertebrates. . Pages 10-27. Freshwater Biomonitoring and Benthic Macroinvertebrates. . Chapman & Hall, New York, USA.

Carolan M. 2006. Science, expertise, and the democratization of the decision-making process. Society & Natural Resources **19**:661-668.

Carr AJL. 2004. Why do we all need community science? Society & Natural Resources **17**:841-849.

Carter JL, Resh VH. 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. Journal of the North American Benthological Society **20**:658-682.

Castilla EP, Cunha DGF, Lee FWF, Loiselle S, Ho KC, Hall C. 2015. Quantification of phytoplankton bloom dynamics by citizen scientists in urban and peri-urban environments. Environmental Monitoring and Assessment **187**.

Church SP, Payne LB, Peel S, Prokopy LS. 2019. Beyond water data: benefits to volunteers and to local water from a citizen science program. Journal of Environmental Planning and Management **62**:306-326.

Clarke RT, Furse MT, Gunn RJM, Winder JM, Wright JF. 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. Freshwater Biology **47**:1735-1751.

Cohn J. 2008. Citizen Science: Can Volunteers Do Real Research? . BioScience:192-197.

Commission for Environmental Cooperation. Undated. Water Quality. The North American Mosaic: An Overview of Key Environmental Issues:4pp.

Connors JP, Lei SF, Kelly M. 2012. Citizen Science in the Age of Neogeography: Utilizing Volunteered Geographic Information for Environmental Monitoring. Annals of the Association of American Geographers **102**:1267-1289.

Conrad CC, Hilchey KG. 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. Environmental Monitoring and Assessment **176**:273-291.

Conrad CT, Daoust T. 2008. Community-based monitoring frameworks: Increasing the effectiveness of environmental stewardship. Environmental Management **41**:358-366.

Cooper C, Hochachka W, Dhondt A. 2012. The Opportunities and Challenges of Citizen Science as a Tool for Ecological Research. Pages 99-113 in Dickinson J, and Bonney R, editors. Citizen Science: Public Participation in Environmental Research. Comstock Publishing Associates, Ithaca and London.

CVRPP. 2017. High Speed 2: Additional Mitigation Plan for the Colne Valley.148 pp.

Davies A. 2001. The use and limits of various methods of sampling and interpretation of benthic macro-invertebrates. Journal of Limnology **60**:1-6.

Dawson M, Hutchins M, Bachiller-Jareno N, Loiselle S. 2019. The spatial and temporal variation of water quality at a community garden site in an urban setting: citizen science in action Freshwater Science **38**:352-364.

Di Fiore D, Fitch B. 2016. The riverfly monitoring initiative: structured community data gathering informing statutory response Environmental SCIENTIST **25**:36-41.

Dickinson JL, Shirk J, Bonter D, Bonney R, Crain RL, Martin J, Phillips T, Purcell K. 2012. The current state of citizen science as a tool for ecological research and public engagement. Frontiers in Ecology and the Environment **10**:291-297.

Dickinson JL, Zuckerberg B, Bonter DN. 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annual Review of Ecology, Evolution, and Systematics, Vol 41 **41**:149-172.

Dunkley RA. 2019. Monitoring ecological change in UK woodlands and rivers: Anexploration of the relational geographies of citizen science. Transactions of the Institute of British Geographers **44**:16-31.
Edwards PM. 2016. The Value of Long-Term Stream Invertebrate Data Collected by Citizen Scientists. Plos One **11**.
EEA. 2015. The European environment — state and outlook 2015: synthesis report.212pp.
Environment Agency. 2016. Annual Fisheries Report 2015 to 2016.100pp.
Environment Agency. 2018. The state of the environment: water quality.12pp.
Environment Agency. 2019. Annual Fisheries Report 2017 to 2018.81pp.
Ettinger W. 1984. Variation Between Technicians Sorting Benthic Macroinvertebrate Samples. Freshwater Invertebrate Biology **3**:147-149.
Firehock K, West J. 1995. A brief-history of volunteer biological monitoring using macroinvertebrates. Journal of the North American Benthological Society **14**:197-202.
Fitch B. 2017. Riverfly Partnership Newsletter, Volume 4, Issue 2, 2017 in Riverfly Partnership, editor.
Fitch B, Brooks S, Brierley B. 2018. A review of the Anglers' Riverfly Monitoring Initiative in England, 2017 to 2018 43pp.
Fitzmaurice G, Laird N. 2015. Linear Mixed Models. International Encyclopedia of the Social & Behavioral Sciences **14**:162-168.
Forrest SA, Holman L, Murphy M, Vermaire JC. 2019. Citizen science sampling programs as a technique for monitoring microplastic pollution: results, lessons learned and recommendations for working with volunteers for monitoring plastic pollution in freshwater ecosystems. Environmental Monitoring and Assessment **191**.
Franca JS, Solar R, Hughes RM, Callisto M. 2019. Student monitoring of the ecological quality of neotropical urban streams. Ambio **48**:867-878.
Friberg N, Sandin L, Furse MT, Larsen SE, Clarke RT, Haase P. 2006. Comparison of macroinvertebrate sampling methods in Europe. Hydrobiologia **566**:365-378.
Furse MT, Wright JF, Armitage PD, Moss D. 1981. An appraisal of pond-net samples for biological monitoring of lotic macro-invertebrates. Water Research **15**:679-689.
Gouveia C, Fonseca A, Camara A, Ferreira F. 2004. Promoting the use of environmental data collected by concerned citizens through information and communication technologies. Journal of Environmental Management **71**:135-154.
Haase P, Lohse S, Pauls S, Schindehutte K, Sundermann A, Rolauffs P, Hering D. 2004a. Assessing streams in Germany with benthic invertebrates: development of a practical standardised protocol for macro invertebrate sampling and sorting. Limnologica **34**:349-365.
Haase P, Pauls S, Sundermann A, Zenker A. 2004b. Testing different sorting techniques in macro invertebrate samples from running waters. Limnologica **34**:366-378.
Hertfordshire Life. 2010. Discover the River Ver in Hertfordshire, https://www.hertfordshirelife.co.uk/out-about/walks/discover-the-river-ver-in-hertfordshire-1-1713238.
Higgins S, Thomas F, Goldsmith B, Brooks S, Hassall C, Harlow J, Stone D, Völker S, White P. 2019. Urban freshwaters, biodiversity, and human health and well-being: Setting an interdisciplinary research agenda. WIREs Water:13pp.

Huddart JEA, Thompson MSA, Woodward G, Brooks SJ. 2016. Citizen science: from detecting pollution to evaluating ecological restoration. Wiley Interdisciplinary Reviews-Water **3**:287-300.

IBM Corp. 2017. IBM SPSS Statistics for Windows. Armonk, NY.

Isaacs A. 2017. Why do volunteers participate in water quality monitoring? Motivations of citizen scientists in the Anglers' Riverfly Monitoring Initiative. Thesis submitted for the degree of MSc Aquatic Science, Dept of Geography, UCL (University College London) 65pp.

Kerans BL, Karr JR, Ahlstedt SA. 1992. Aquatic invertebrate assemblages- spatial and temporal differences among sampling protocols. Journal of the North American Benthological Society **11**:377-390.

Kiessling T, Knickmeier K, Kruse K, Brennecke D, Nauendorf A, Thiel M. 2019. Plastic Pirates sample litter at rivers in Germany - Riverside litter and litter sources estimated by schoolchildren. Environmental Pollution **245**:545-557.

Koo TK, Li MY. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine **15**:155-163.

Kripa P, Prasanth K, Sreejesh K, Thomas T. 2012. Aquatic Macroinvertebrates as Bioindicators of Stream Water Quality- A Case Study in Koratty, Kerala, India. Research Journal of Recent Sciences **2**:217-222.

Latimore JA, Steen PJ. 2014. Integrating freshwater science and local management through volunteer monitoring partnerships: the Michigan Clean Water Corps. Freshwater Science **33**:686-692.

Letovsky E, Myers I, Canepa A, McCabe D. 2012. Differences between kick sampling techniques and shortterm Hester-Dendy sampling for stream macroinvertebrates **83**:47-55.

Levesque D, Cattaneo A, Deschamps G, Hudon C. 2017. In the eye of the beholder: Assessing the water quality of shoreline parks around the Island of Montreal through citizen science. Science of the Total Environment **579**:978-988.

Li L, Zheng BH, Liu LS. 2010. Biomonitoring and Bioindicators Used for River Ecosystems: Definitions, Approaches and Trends. International Conference on Ecological Informatics and Ecosystem Conservation (Iseis 2010) **2**:1510-1524.

Louw M, Muenz T, Roberts J, Wilson M, Kerlin S. 2018. Aquatic Macroinvertebrate Identification Trainings for Volunteers: Results of a National Materials and Practices Inventory Survey. Technical report to be published through the Carnegie Mellon University School of Computer Science:56pp.

Macadam C. 2016. A review of the status of the mayflies (Ephemeroptera) of Great Britain - Species Status No.28. Natural England Commissioned Reports, Number 193:54pp.

Mackey A, Cooling D, Berrie A. 1984. An evaluation of sampling strategies for qualitative surveys of macroinvertebrates in rivers, using pond nets Journal of Applied Ecology **21**:515-534.

Maxwell S, Delaney H, Kelley K. 2017. Designing Experiments and Analyzing Data: A Model Comparison Perspective.1071pp.

Mellanby K. 1974. A water pollution survey, mainly by British children. Environmental Pollution **6**:161-173.

Metcalfe JL. 1989. Biological water-quality assessment of running waters based on macroinvertebrate communities- history and present status in Europe. Environmental Pollution **60**:101-139.

Miguel-Chinchilla L, Heasley E, Loiselle S, Thornhill I. 2019. Local and landscape influences on turbidity in urban streams: a global approach using citizen scientists **38**:303-320.
Mkandla A. 2018. HNL Appraisal Package 2 Pinn and Cannon Brook Initial Assessment Plus Document.32pp.
Moffett ER, Neale MW. 2015. Volunteer and professional macroinvertebrate monitoring provide concordant assessments of stream health. New Zealand Journal of Marine and Freshwater Research **49**:366-375.
Muirhead S. 2011. Nature and well-being : building social and emotional capital through environmental volunteering. University of Dundee.
National Rivers Authority. Undated. Fact File- River Chess and Misbourne.4pp.
National Riverwatch. 1994. The River Report. A three year project review.
Oti IC, Gharaibeh NG, Hendricks MD, Meyer MA, Van Zandt S, Masterson J, Horney JA, Berke P. 2019. Validity and Reliability of Drainage Infrastructure Monitoring Data Obtained from Citizen Scientists. Journal of Infrastructure Systems **25**.
Owen R, Parker A. 2018. Citizen science in environmental protection agenceis. Page 582 in Hecker S, Haklay M, Bowser A, Makuch Z, Vogel J, and Bonn A, editors. Citizen science: innovation in open science, society and policy. UCL Press, London, UK.
Paisley MF, Trigg DJ, Walley WJ. 2014. Revision of the biological monitoring working party (BMWP) score system: derivation of present-only and abundance-related scores from field data. River Research and Applications **30**:887-904.
Pillemer K, Fuller-Rowell TE, Reid MC, Wells NM. 2010. Environmental Volunteering and Health Outcomes over a 20-Year Period. Gerontologist **50**:594-602.
Pocock M, Chapman D, Sheppard L, Roy H. 2014. Choosing and Using Citizen Science: a guide to when and how to use citizen science to monitor biodiversity and the environment.:28pp.
Pollock R, Whitelaw G. 2005. Community-Based Monitoring in Support of Local Sustainability. Local Environment **10**:211-218.
Radlett Neighbourhood Plan Steering Group. 2019. The Radlett Plan.78pp.
Ramos-Merchante A, Prenda J. 2017. Macroinvertebrate taxa richness uncertainty and kick sampling in the establishment of Mediterranean rivers ecological status. Ecological Indicators **72**:1-12.
Ratnieks FLW, Schrell F, Sheppard RC, Brown E, Bristow OE, Garbuzov M. 2016. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods in Ecology and Evolution **7**:1226-1235.
Riverfly Partnership. 2007. Company fined for killing fish and polluting Sirhowy River Environment Agency 26 September 2007, http://www.riverflies.org/press-releases. (accessed 7/7/2019.
Rose NL, Turner SD, Goldsmith B, Gosling L, Davidson TA. 2016. Quality control in public participation assessments of water quality: the OPAL Water Survey. Bmc Ecology **16**.
Roy H, Pocock M, Preston C, Roy D, Savage J, Tweddle J, Robinson L. 2012. Understanding Citizen Science and Environmental Monitoring 175pp.
Royle JA. 2004. Modeling abundance index data from anuran calling surveys. Conservation Biology **18**:1378-1385.
Seelen LMS, et al. 2019. An affordable and reliable assessment of aquatic decomposition: Tailoring the Tea Bag Index to surface waters. Water Research **151**:31-43.

Silvertown J. 2009. A new dawn for citizen science. Trends in Ecology & Evolution **24**:467-471.
STAR. 2004. UK Invertebrate sampling and analysis procedure for STAR project. Page 63pp.
Stepenuck KF, Genskow KD. 2019. Traits of Volunteer Water Monitoring Programs that Influence Natural Resource Management and Policy Impacts. Society & Natural Resources **32**:275-291.
Storey RG, Wright-Stow A, Kin E, Davies-Colley RJ, Stott R. 2016. Volunteer stream monitoring: Do the data quality and monitoring experience support increased community involvement in freshwater decision making? Ecology and Society **21**.
Sultana P, Abeyasekera S. 2008. Effectiveness of participatory planning for community management of fisheries in Bangladesh. Journal of Environmental Management **86**:201-213.
The Riverfly Partnership. 2017. ARMI Training Presentation.
Thompson MSA, et al. 2016. Gene-to-ecosystem impacts of a catastrophic pesticide spill: testing a multilevel bioassessment approach in a river ecosystem. Freshwater Biology **61**:2037-2050.
Thornhill I, Chautard A, Loiselle S. 2018. Monitoring Biological and Chemical Trends in Temperate Still Waters Using Citizen Science. Water **10**.
Thornhill I, Loiselle S, Clymans W, Noordwijk C. 2019. How citizen scientists can enrich freshwater science as contributors, collaborators, and co-creators Freshwater Science **38**:231-235.
Tredick CA, Lewison RL, Deutschman DH, Hunt TA, Gordon KL, Von Hendy P. 2017. A Rubric to Evaluate Citizen-Science Programs for Long-Term Ecological Monitoring. Bioscience **67**:834-844.
Turner SD, Rose NL, Goldsmith B, Bearcock JM, Scheib C, Yang H. 2017. Using public participation to sample trace metals in lake surface sediments: the OPAL Metals Survey. Environmental Monitoring and Assessment **189**.
UN Water. 2016. Towards a Worldwide Assessment of Freshwater Quality: A UN-Water Analytical Brief 40pp.
UNEP. 2016. Hundreds of Millions Face Health Risk as Water Pollution Rises Across Three Continents, https://www.unenvironment.org/news-and-stories/story/hundreds-millions-face-health-risk-water-pollution-rises-across-three. (accessed 25th August 2019.
Vlek HE, Sporka F, Krno I. 2006. Influence of macroinvertebrate sample size on bioassessment of stream. Hydrobiologia **566**:523-542.
Wallace I. 1991. A review of the Trichoptera of Great Britain. Research and Survey in Nature Conservation **32**:1-59.
Wallace I. 2003. The Beginner's Guide to Caddis (Order Trichoptera). Bulletin of the Amateur Entomologists' Society **62**:132pp.
Waterton C. 2003. Messing about on the river. Pages 56-58. Salmo trutta.
Weigelhofer G, Pölz E, Hein T. 2019. Citizen science: how high school students can provide scientifically sound data in biogeochemical experiments Freshwater Science **38**:236-243.

8. GIS Data Sources

Figure 2: OS Data (2019), accessible at: https://www.ordnancesurvey.co.uk/business-government/products