

Evaluation of drinking water quality test kits for home use in the United States

LeighAnn D'Andrea^{a,b} and Emily Kumpel^{IWA id a,*}

^a Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Road, Amherst, MA 01003, USA

^b United States Air Force, 3465 North Carolina Ave, Upper Marlboro, MD 20762, USA

*Corresponding author. E-mail: ekumpel@umass.edu

 EK, 0000-0003-0138-8441

ABSTRACT

Water consumers in the United States may want to test their drinking water using at-home commercially available test kits rather than a certified laboratory due to convenience and affordability. However, while numerous do-it-yourself test kits are available for purchase online or at local stores, these kits are unregulated and lack data on their performance. We evaluated off-the-shelf home drinking water test kits that measure iron, copper, manganese, and fluoride concentrations to investigate whether these kits could reliably provide meaningful results. We evaluated their performance in three water matrices: deionized water (DI), tap water, and river water, and with laboratory-trained personnel compared to untrained users. Our results showed highly repeatable but variable performance in the test kits' ability to detect potential contaminants in the water. Most kits performed best in the DI water matrix with no interference. Our results suggest that there are concerns about their accuracy and usefulness and that whether the results can be relied on depends on which parameter is being measured in which water with which kit and for which purpose.

Key words: citizen science, drinking water, field testing methods, water quality testing

HIGHLIGHTS

- Off-the-shelf water quality test kits for measuring iron, copper, manganese, and fluoride were evaluated.
- Test kits showed variable performance.
- Many test kits performed well in deionized water but performed poorly when measuring concentrations in tap or river water.
- While many test kits were not accurate, they were still able to inform potential users of above or below regulatory limits.

INTRODUCTION

While the United States reports high access to safely managed drinking water, recent analyses have highlighted disparities in communities' access to safe drinking water. While the majority of the population in the United States is served by a piped public water system, which is regulated under the Safe Drinking Water Act, another 20 million houses in the United States are estimated to be served by private water sources, such as wells, which are largely unregulated (National Research Council 1997; US EPA 2015a; Murray *et al.* 2021). Even among public water systems that are monitored regularly for water quality, not every service connection is tested, and water from premise plumbing is largely excluded from sampling (except for some parameters regulated at the tap, such as lead and copper). Lack of trust in tap water is reported across the United States and has been recognized as a public health concern (Patel & Schmidt 2017; Pierce & Gonzalez 2017; Pierce *et al.* 2019). Consumers of unregulated sources such as private wells often want to test their well water by sending samples to a private lab, participating in a program that facilitates testing, or buying home test kits (Flanagan *et al.* 2015).

State agencies recommend that if a homeowner wants to test their drinking water – whether from a public water supplier or a private well – they use a state-certified lab, with fees set by the individual labs. However, there exist numerous do-it-yourself test kits available for private individuals to purchase online or at local stores that consumers looking for a more affordable option may elect to do. These do-it-yourself kits that are commercially sold do not undergo any formal certification or accreditation process to ensure their accuracy. Prior studies have evaluated the ability of off-the-shelf (OTS) test kits, including test strips and colorimetric vials, to accurately and precisely measure lead (Kriss *et al.* 2021), chlorine residual (Murray &

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

Lantagne 2015), arsenic (George *et al.* 2012; Powers *et al.* 2019; Reddy *et al.* 2020), and nitrate (Nielsen *et al.* 2008; Aukema & Wackett 2019), among others. Overall, there has been wide variability reported in the ability of these test kits to measure these water quality parameters accurately or in their ability to accurately classify a water sample as above or below a threshold (whether that be of detection or above a regulatory limit or guideline). However, to date, evaluations of test kits for home users have been reported in the literature for only a limited set of water quality parameters (lead, nitrate, and arsenic; Nielsen *et al.* 2008; Reddy *et al.* 2020; Kriss *et al.* 2021), and often the test kits evaluated have been single-parameter methods, while much of what is available on the market and marketed for consumers are multi-parameter tests. While these test kits are widely available, consumers who already do not know the status of their drinking water and may lack knowledge of water quality also do not know whether to trust the results of these test kits. In particular, consumers may want to measure water quality parameters that may originate from the distribution system and premise plumbing systems; for example, being able to measure and understand the source of discolored water, which often originates from elevated iron or manganese, may not be a health risk but is an esthetic concern important for water consumers (Tang *et al.* 2018; Vidmar *et al.* 2023).

In this study, we evaluated whether available OTS test kits could accurately measure several water quality parameters (copper, iron, fluoride, and manganese) in different water matrices and by non-specialist users to identify how well currently available methods perform.

METHODS

Water matrices

We performed experiments using three water matrices: deionized (DI) water, tap water, and river water. DI water was used to represent a control of high purity with no background interference. Tap water was used to represent water that consumers may test; we obtained the tap water from a tap at the University of Massachusetts Amherst campus, supplied by the Amherst, MA, public water system, which supplies treated surface water. River water was obtained from the Mill River (Amherst, MA), a nearby untreated source of water. We sought to represent different water matrices by including both treated and untreated surface water to evaluate the impact of organic matter and other ions on measurement. We originally intended to use a phosphate buffer to hold the pH of each solution at high and low pHs but the addition of the buffer interfered with the iron solutions. Instead, we recorded the pH and temperature for each solution but did not adjust either.

Parameters for analysis

We selected four drinking water constituents for analysis: iron, copper, manganese, and fluoride. In the United States, these are regulated by the US EPA under the Safe Drinking Water Act as primary or secondary standards (maximum contaminant level or secondary maximum contaminant level (SMCLs)) (US EPA 2015b). Iron has an SMCL of <0.3 mg/L; while concentrations above the SMCL do not pose a health risk but can negatively affect the esthetics or taste of water and cause infrastructure damage. Copper has a maximum contaminant level (MCL) of <1.3 mg/L and an SMCL of <1.0 mg/L, as concentrations above this will cause a metallic taste and a blue staining. Copper in drinking water systems is regulated by the EPA under the lead and copper rule, as copper can enter the water due to corrosion in premise plumbing. Manganese also has an SMCL of <0.05 mg/L; at higher concentrations, consumers will notice a black or brown color, black staining, and a bitter taste. Fluoride is both a primary and secondary contaminant with an MCL of <4.0 mg/L and an SMCL of <2.0 mg/L. While no adverse health effects are expected between 2.0 and 4.0 mg/L, prolonged exposure may cause tooth discoloration. Concentrations >4.0 mg/L can cause fluorosis (bone disease) (Srivastava & Flora 2020).

Selection of test kits

We purchased OTS test kits from a major online retailer. While consumers may purchase test kits from local stores (e.g. hardware stores), we sought nationally available kits and therefore purchased from a nationally available retailer (www.amazon.com). Kits were selected based primarily on those that appeared as top-ranked kits in the supplier algorithm at the time of the search, as a consumer might decide to purchase a kit. We selected kits measuring single or multiple ('multiparameter') constituents simultaneously: four multiparameter kits, four iron-only kits, two copper-only kits, and one manganese-only kit (Table 1). There are hundreds of test kits on the market in the United States at any given time, and the availability of what is on the market or in stock can change daily; therefore, we did not seek to conduct a thorough study of all kits available on the market, but rather to use a process of a consumer attempting to identify a test kit to use and following through with

Table 1 | Range and increments of OTS kits

Test kit	Range (mg/L)	Increments
Iron		
M1	0–5	0, 0.3, 0.5, 1, 3, 5, 10, 25, 50, 100
M2	0–500	0, 5, 10, 25, 50, 100, 250, 500
M3	0–500	0, 5, 10, 25, 50, 100, 250, 500
M4	0–500	0, 5, 10, 25, 50, 100, 250, 500
SA1	0–5	0, 0.15, 0.30, 0.60, 1, 2, 5
SB2	0–100	0, 2, 5, 10, 25, 50, 100
SC3	0–5	0, 0.02, 0.05, 0.10, 0.20, 0.30, 0.50, 0.75, 1, 2, 5
SD4	0–5	0, 0.3, 0.5, 1, 3, 5
Copper		
M1	0–10	0, 0.5, 1, 3, 10
M2	0–10	0, 0.5, 1, 3, 10
M3	0–10	0, 0.5, 1, 3, 10
M4	0–300	0, 1, 10, 30, 100, 300
SA5	0–3	0, 0.2, 0.5, 1, 3
SC6	0–2	0, 0.05, 0.1, 0.2, 0.4, 1, 2
Fluoride		
M1	0–5	0, 0.5, 1, 2, 4, 5
M2	0–100	0, 10, 25, 50, 100
M3	0–100	0, 10, 25, 50, 100
M4	0–5	0, 0.5, 1, 2, 5
Manganese		
SC7	0–1.6	<0.02, 0.05, 0.1, 0.2, 0.4, 1.0, > 1.6

Note. M stands for multi-parameter, S indicates single parameter; A, B, C, and D indicate different test kits within that parameter designation (for example, SA1 and SA5 represent the same kit that measured both iron and copper).

whether this kit may provide reliable water quality results. Costs for each test kit were recorded at the time of purchase and retailer (2019) to provide overall context and comparability between selected kits.

Many evaluated multiparameter test kits measured a higher concentration of iron than those measuring only iron (maximum detection limit (MDL) of 500 mg/L) (Table 1). Three of the kits measuring only iron claimed an MDL of 5 mg/L and one of 100 mg/L. While the multiparameter kits measured higher concentrations, they had larger intervals between measurement points (e.g. a multiparameter kit may measure 0, 5, and 10 mg/L of iron, while an iron-only kit measures 0, 0.02, 0.05, 0.1, 0.2, and 0.3 mg/L). Similar features were observed for copper kits (Table 1). No test kit that measured only fluoride was available via the online retailer and therefore not evaluated in this research.

Target and measured concentrations

To select stock solution increments for evaluating test kits, we aligned target concentrations with the relevant MCL/SMCLs and increments that test kits were marketed to measure. These were:

- Iron: 0.1, 0.3, 0.5, 1.0, 3.0, 4.0, 5.0, 10, 25, 50, and 100 mg/L (note: the SMCL is <0.3 mg/L).
- Copper: 0.1, 0.2, 0.4, 1.0, 2.0, and 3.0 mg/L (SMCL of 1.0 mg/L).
- Manganese: <0.02, 0.05, 0.1, 0.4, and 1.0 mg/L (SMCL: 0.05 mg/L).
- Fluoride: 0.50, 1.0, 2.0, 4.0, 10, 25, and 50 mg/L (MCL: 4.0 mg/L; SMCL: 2.0 mg/L)

For each constituent, we prepared a stock solution with high concentrations of the constituents and then diluted it using the appropriate water matrix to achieve a range of lower concentrations. A 10 mg/L stock solution of iron was made by adding

0.0251 g of ferrous sulfate heptahydrate ($\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$) to 500 mL of DI water. To test the higher concentrations measured by some test kits, a second stock solution of 150 mg/L was made by adding 0.375 g of $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ to 500 mL of DI water. For copper, a 10 mg/L stock solution of copper was made by adding 0.0391 g of cupric sulfate (CuSO_4) to 500 mL of DI water. A 2 mg/L stock solution of manganese was made by adding 0.0062 g of manganese sulfate monohydrate ($\text{MnSO}_4 \cdot \text{H}_2\text{O}$) to 500 mL of DI water. Two fluoride stock solutions of 5 and 100 mg/L were made by adding 0.0056 and 0.1105 g of sodium fluoride (NaF), respectively, to 500 mL of DI water. We selected sulfate salts for the addition of iron, copper, and manganese. These anions can affect the behavior of metal in the water, with implications for measurement methods; however, we sought to add in compounds that were highly soluble in water, had well-characterized dissociation behavior, and represented those naturally present in drinking water, particularly as related to distribution systems or household plumbing. NaF can be used for water fluoridation in drinking water systems and is used as a calibration solution for the measurement of fluoride with ion-selective electrodes (ANSI/AWWA 2011; US EPA 2015).

The target test concentrations were then made using dilutions of the stock solutions (Supplementary material, Table S1). For tap and river water, adjustments were made for background concentrations already present in the water sample. For example, during one experiment day, the measured iron concentration in the Amherst tap water was 0.04 mg/L; therefore, the stock solution added to the Amherst tap water was adjusted to meet the desired end concentration (e.g. 0.3 mg/L). The goal was to keep the target concentrations uniform across the three water matrix experiments. The target concentrations and measured concentrations were recorded (Table S1). All solutions were made fresh immediately before experiments.

We tested the solution concentrations with a Hach DR2700 spectrophotometer, which we refer to as the 'laboratory-obtained concentration,' following the instructions for each method. While other analytical methods (e.g. inductively coupled plasma mass spectrometry (ICP-MS)) would have yielded more accurate results for our true value, we selected a spectrophotometer-based method as a comparison based on colorimetric processes used by the test strip approaches, as what many water utilities would use as field test kits. Each used a 10 mL sample vial, and some had to be diluted depending on the MDL. A DI blank was used for all methods. Iron was measured with FerroVer Iron Reagent Power Pillows from Hach following Method 8008, adapted from the Standard Methods for the Examination of Water and Wastewater (range: 0.02–3.00 mg/L). This method converts all soluble and most insoluble iron to soluble ferrous iron and specifies that copper concentrations will not interfere due to the presence of a masking agent included in the reagent. Copper was measured with CuVer Copper Reagent Power Pillows from Hach following Method 8506, the US EPA Bicinchoninate Method (detection range: 0.04–5.0 mg/L). This method 8506 reduces Cu^{2+} to Cu^+ and then the bicinchoninate reacts with the Cu^+ to form a purple-colored complex. Manganese was measured following Hach's Manganese LR PAN Method (Method 8149) (detection range: 0.006–0.70 mg/L). This method uses ascorbic acid to reduce all oxidized forms of manganese to Mn^{2+} and an alkaline cyanide reagent to mask any potential interferences and a PAN indicator that forms an orange-colored complex with Mn^{2+} . Fluoride was measured using Hach's USEPA SPADNS 2 Method (Method 10225) (detection range: 0.02–2.00 mg/L). This method involves fluoride reacting with a red zirconium dye (a SPADNS 2 reagent), wherein the fluoride combines with parts of the zirconium to form a colorless complex that bleaches the red color in an amount proportional to the fluoride concentration.

Experimental procedure

We measured each solution's concentration using the spectrophotometer (Hach, Loveland, CO) and the appropriate method. We filled 250 mL glass beakers labeled by the target concentration of that parameter with the solution. We recorded the pH and temperature solution (from lowest to highest). For the DI water matrix, only concentration and temperature were recorded. To measure the OTS kits, the solutions remained in the 250 mL glass beakers, and the instructions were followed for each kit. Some methods included transferring the solution from the beaker into a provided sample vial; otherwise, the sample was transferred from the glass beaker to a 100 or 250 mL plastic cup. Each test kit brand was used five times on each solution, following that kit's specific instructions and recording the result after waiting the specified time. Each solution was tested with each brand's test in quintuplicate.

Usability testing

We recruited 29 participants from undergraduate student groups from the University of Massachusetts Amherst College of Engineering to use the test kits as representatives of potential citizen scientists ('untrained users') in April 2019. The study was reviewed and determined to be exempt by the University of Massachusetts Amherst Institutional Review Board (Protocol

ID 2019-5435). We selected one multiparameter kit and two iron-only test kits to test in two water sources (DI water and tap water). We selected the kits that performed the best (most accurate) in the laboratory tests. Just before the meetings, test solutions were freshly made. The laboratory-obtained values for the concentrations of these solutions (using the Hach methods) were recorded, and then the solutions were put into color-coded containers. A data sheet was given to the students to record their results individually and in groups of two or more. A survey was also included on this data sheet to get input from these untrained users on their experience using the test kits.

Statistical analysis

Calculations and figures were made using R (R Core Team 2024). Sensitivity was calculated for each test kit for each source of water to assess its ability to provide accurate results based on binary measures of (1) ≥ 0 mg/L (i.e. whether, if there was any contaminant present in the water, the test kits were able to measure anything > 0 mg/L); or (2) \geq the EPA MCL or SMCL (Figure 1). Sensitivity was calculated as the number of true positives divided by the sum of the number of true positives and false negatives (Equation (1)). Under the first definition, a true positive was defined as a test strip reporting any detection (Table 2) while a false negative was defined as a test strip reporting a concentration of 0 mg/L when the actual concentration was > 0 mg/L. For the second definition, a true positive was defined as the strip reporting a concentration greater than or equal to that constituent's MCL/SMCL and the actual concentration being \geq that constituent's MCL/SMCL (Table 2). Therefore, a false negative was when the test strip reported a concentration less than the SMCL of a constituent, but the actual

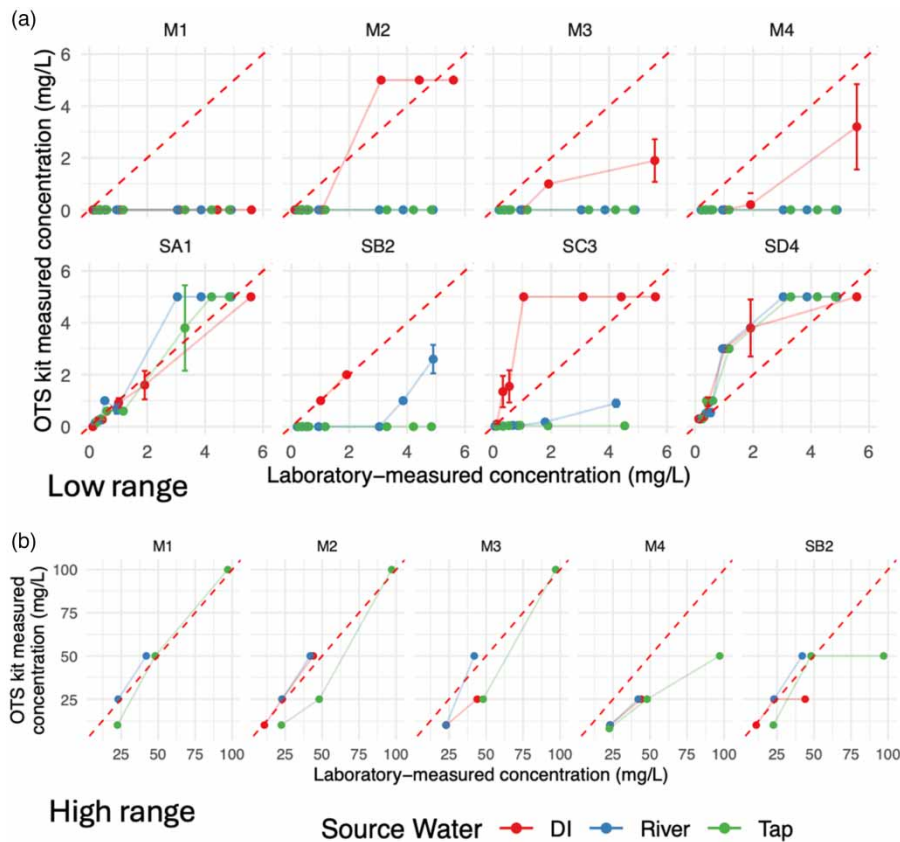


Figure 1 | Iron was measured at (a) lower ranges (0–6 mg/L) and (b) higher ranges (6–100 mg/L) as measured by test kits compared to the laboratory-obtained concentrations (FerroVer Iron Reagent). Each row represents a different kit and each column a different source of water. The diagonal lines represent a perfect match between FerroVer Iron Reagent Power Pillow measurements (taken as the laboratory-obtained accurate concentration) and the test kit. Each data point represents the mean of five trials per solution and a 95% confidence interval; if no error bars are visible, all five tests yield the same results. The top row represents multiparameter kits, while the bottom row represents iron-only test kits.

Table 2 | Definitions for sensitivity calculation based on detection (0 mg/L or >0 mg/L) or the USEPA's MCL/SMCLs (>0 mg/L or ≥MCL/SMCL)

	Spec = 0 or < MCL/SMCL	Spec >0 or ≥ MCL/SMCL
Strip = 0 or <MCL/SMCL	True negative	False negative
Strip >0 or ≥MCL/SMCL	False positive	True positive

concentration was greater than the SMCL:

$$\text{Sensitivity} = \frac{\# \text{ true positive}}{\# \text{ true positive} + \# \text{ false negative}} \times 100 \quad (1)$$

RESULTS

Precision and accuracy

The majority of test kit-concentration measurements had a standard deviation of zero, meaning that the five replicate readings were the same (see [Figure 1](#) with points that did not have error bars, meaning the SD = 0). The control with DI water (nothing added) showed that test kits could measure 0 mg/L. Therefore, the test results were typically precise (consistent between repeated tests).

Iron

All multiparameter test kits were able to accurately measure when iron was 0 mg/L in the DI water ([Figure 1](#)). Many of them were able to detect iron in the DI water matrix but not in river or tap water ([Figure 1](#)). One multi-parameter kit (M1) could not detect iron in any source waters, except one sample in river water at 7.5 mg/L (although it underestimated this concentration by 5 mg/L). Another (M2) could not detect iron in the DI water until 3.10 mg/L (estimating it as 5 mg/L), after which points were closer to the match line. However, this kit did not detect any iron in the tap water and underestimated concentrations in river water. The two other multi-parameter kits (M3 and M4) underestimated but detected iron in the DI water but were unable to measure iron of any concentration in tap or river water. These kits were also the only ones with any points with an SD > 0.

The iron-only kits could all accurately measure 0 mg/L in the DI water ([Figure 1](#)). One single-parameter kit (SA1) measured concentrations close to the laboratory-obtained concentrations, while, in river water, it slightly over or underestimated concentrations until >2 mg/L. Another single-parameter kit (SB2) closely matched the laboratory-obtained concentration when in the DI water (at 1.01, 1.91, and 5.58 mg/L, this kit gave the results 1, 2, and 6 mg/L, respectively) however, in tap water, it did not detect any iron until a concentration of 7.6 mg/L (which it estimated as 2 mg/L). In river water, the same kit underestimated all concentrations as 0 mg/L until 3.86 mg/L, which it estimated as 1 mg/L. The SC3 had primarily overestimated in DI water and underestimated in tap and river water, and the final single-parameter kit (SD4) overestimated concentrations in all source waters.

Several kits were advertised to measure high iron concentrations (20–100 mg/L). In tap and river water, one multiparameter kit (M1) was able to measure some concentrations resembling the laboratory-obtained results (e.g. targets of 22.5, 48, and 97 mg/L were estimated as 10, 50, and 100 mg/L, respectively, in tap water), performing better at 20–100 mg/L than at 0–12 mg/L ([Figure 1](#)). Another single-parameter kit (SB2) performed reasonably well in tap and river water based on kit increments or underestimated higher concentrations (the data points are not visible in [Figure 2](#)) (e.g. targets 48 and 97 mg/L as 50 mg/L in tap water). In DI water, one kit (M2) provided very close estimates, but in tap water, it underestimated target concentrations. Kit M3 was close to the actual in DI water (concentrations of 23, 44, and 107 mg/L measured by the kit as 10, 25, and 100 mg/L, respectively), but underestimated results in tap water (22.5 and 48 mg/L as 5 and 25 mg/L, respectively) while correctly measuring high concentrations (97 mg/L as 100 mg/L). This same kit performed more variably in river water ([Figure 1](#)). The M4 kit underestimated most concentrations in all water matrices. Overall, the multiparameter kits performed better at higher concentrations than at lower concentrations ([Figure 1](#)). Notably, as the concentration of iron

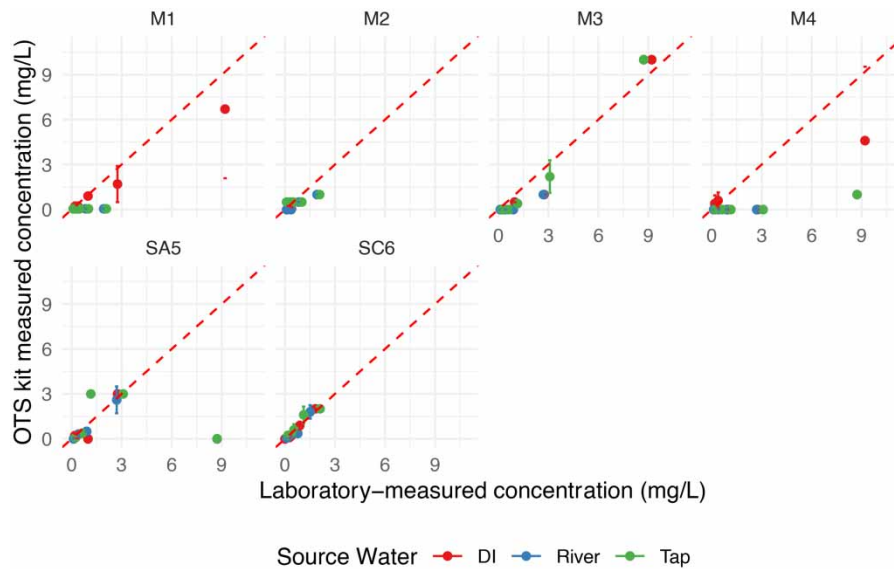


Figure 2 | Copper concentrations (0–12 mg/L) as measured by a multiparameter test kits (top row) and copper-only test kits (bottom row) compared to the laboratory-obtained concentration (CuVer Iron Reagent).

increased, the pH decreased (Figure S1). pH will affect the solubility of iron and therefore is expected to affect the form of iron in the water (and kit detection).

Copper

Among the multiparameter kits, one (M1) was able to detect copper in DI water but consistently underestimated concentrations in tap and river water, with larger departures at higher concentrations (Figure 2). Similarly, another kit (M2) measured 0.19 and 0.39 mg/L as 0.1 and 0.05 mg/L, respectively, in DI water, but in tap water measured all concentrations as 0.05 mg/L until 2.08 mg/L, which it estimated as 1 mg/L (with similar trends in river water). Another kit (M3) provided better measurements at higher concentrations than lower concentrations in all water types. The final kit (M4) estimated all concentrations in tap and river water as 0 mg/L until 8.72 mg/L, which it underestimated as 1 mg/L. For the copper-only tests, the SC6 performed better than the SA5 at low concentrations, the latter of which yielded a mix of over- and underestimates in the DI water and in tap water, and underestimated concentrations in river water. Overall, the SC6 kit had a data point close to the perfect match line across all three water sources.

Fluoride

Four test kits (all multiparameter kits) were tested for their ability to measure fluoride concentrations 0–12 mg/L (Figure 3). Most kits could not detect fluoride at all, and those that did were underestimates.

Manganese

One test kit (SC7) was analyzed for its ability to measure manganese solutions, which was able to obtain estimates close to accurate concentrations in all three water sources (Figure 4).

Sensitivity

Iron

The top four best-performing kits were those that measured only iron, and the four worst-performing kits were the multiparameter kits (Table 3). The SD4 kit had the highest sensitivity in each water source compared to all other kits, with 0.86, 1, and 1 in the DI, river, and tap water, and ranked first and second out of all the kits according to both sensitivity measures in both tap water and river water for the ability to detect at the SMCL. The SA1 kit also performed well and was more sensitive in tap and river waters than DI water. The SB2 kit performed much better at determining whether

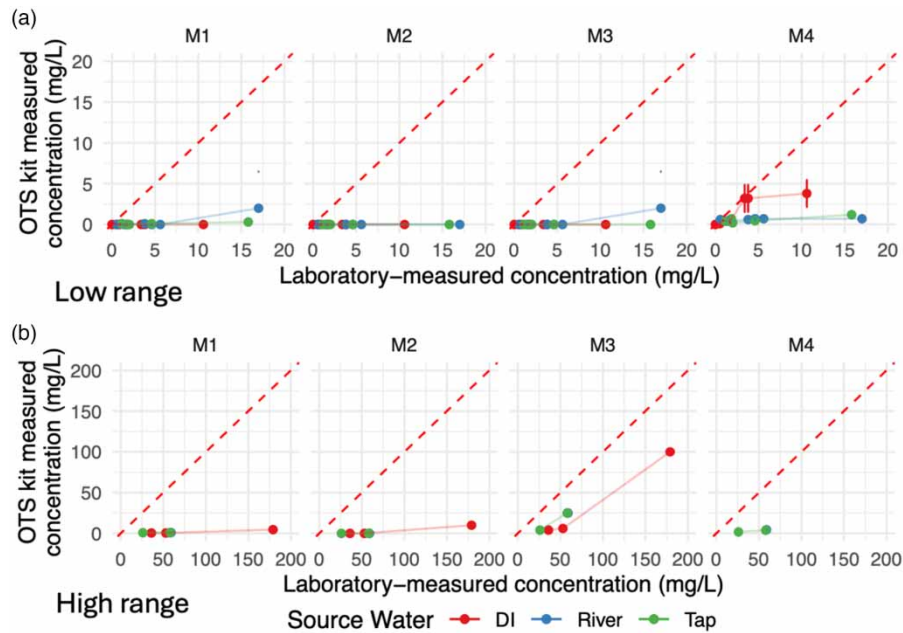


Figure 3 | Fluoride concentrations as measured by multiparameter test kit compared to the laboratory-obtained concentration (SPADNS 2 method) x-axis shown from (a) 0–20 mg/L and (b) 0–200 mg/L.

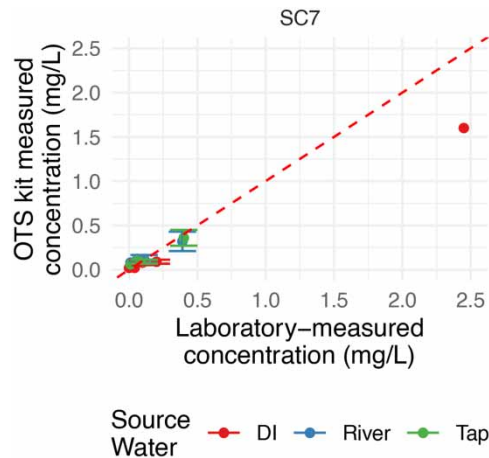


Figure 4 | Manganese concentrations (0–2.5 mg/L) as measured by the manganese-only test kit compared to the laboratory-obtained concentration (PAN LR Method); one value is not shown due to the scale where the laboratory-obtained value was 7.70 mg/L and the mean measured test kit value was 1.60 mg/L (SD = 0).

a value was above or below the SMCL than in its ability to detect. Notably, the SC3 kit performed well in its ability to detect but not when detecting the SMCL.

Copper

As with iron, the SC6 kit performed well for detection but performed worse at SMCL detection. Some of the multiparameter kits outperformed the kits that measured only copper concentration. The M1 kit outperformed SA5 in detection but performed the worst of all the kits for SMCL detection. Notably, the highest sensitivity value out of any of the kits based on the SMCL sensitivity analysis was only 0.40, which is very low; most kits seemed to be designed for detection, but not at the level of the SMCL.

Table 3 | Sensitivity analysis of kits

	Ability to detect				SMCL			
	DI Water	Tap Water	River Water	Total	DI Water	Tap Water	River Water	Total
Iron								
SD4	0.86	1.00	1.00	0.95 ^a	0.71	0.86	0.86	0.81 ^a
SA1	0.71	1.00	1.00	0.90 ^{**}	0.63	0.86	0.80	0.76 ^{***}
SC3	1.00	0.86	0.83	0.90 ^{**}	0.86	0.00	0.14	0.33
SB2	1.00	0.57	0.86	0.81	1.00	0.57	0.86	0.81 ^a
M2	1.00	0.43	0.57	0.67	1.00	0.43	0.57	0.67
M3	0.86	0.43	0.43	0.57	0.86	0.43	0.43	0.57
M4	0.74	0.43	0.43	0.53	0.74	0.43	0.43	0.53
M1	0.00	0.43	0.57	0.33	0.00	0.43	0.57	0.33
Copper								
SC6	1.00	0.80	1.00	0.93 ^a	0.20	0.40	0.20	0.27 ^{**}
M1	0.76	1.00	1.00	0.92 ^{**}	0.32	0.00	0.00	0.11
SA5	0.52	0.68	0.60	0.60 ^{***}	0.20	0.40	0.20	0.27 ^{**}
M3	0.60	0.56	0.40	0.52	0.40	0.40	0.40	0.40 ^a
M2	0.12	1.00	0.40	0.51	0.00	0.20	0.20	0.13
M4	0.40	0.20	0.20	0.27	0.20	0.20	0.20	0.20
Fluoride								
M4	0.56	0.63	0.80	0.66 ^a	0.33	0.22	0.20	0.25 ^a
M1	0.33	0.33	0.27	0.31 ^{**}	0.11	0.00	0.02	0.04
M3	0.22	0.16	0.18	0.19	0.22	0.16	0.18	0.19 ^{**}
M2	0.11	0.00	0.00	0.04	0.11	0.00	0.00	0.04
Manganese								
SC7	0.80	1.00	1.00	0.93	0.60	0.80	0.80	0.73

Note. Each row represents a different test kit and each column represents a different water matrix. The total column takes the average of the three water matrix values. ^aFirst place; ^bsecond place; ^cthird place.

Fluoride

The sensitivity analysis found that kits performed better on their ability to detect rather than on their ability to detect at the SMCL. The M4 kit performed the best and the M2 kit the worst in both analyses. The kits typically performed better in the DI water than other water types.

Manganese

The SC7 kit had a higher sensitivity for detection than the SMCL and performed better in detection in tap and river water than in DI water. While there were no other kits on the market available for comparison, these sensitivity values (0.93 and 0.73) were high.

Costs and instructions

The multiparameter kits capable of testing multiple parameters cost between \$0.13 and \$0.19/test (with each test evaluating for 10–14 parameters within ‘one test’); all of these cost less than the single parameter-only test (Table 4). The iron-only tests ranged from \$0.38 to \$1.01/test (with ‘one test’ testing a sample for only iron), while the copper-only test kits were \$0.52–\$0.53/test (for copper only), and the manganese-only at \$1.18 per test.

Only four of the 11 evaluated test kits provided guidance on the amount of water to use for each test, and none provided guidance on whether or how to flush or not flush (i.e. let water run through pipes for a specified amount of time or the time of day to take the water sample (Table 4).

All four multiparameter kits instructed users to dip a strip into a water sample for a set time, remove it while shaking off excess water, and then wait 15–60 s (kit-dependent) before asking the user to match the color on the strip to the chart on the test kit bottle. One kit (the M3) specified how long to wait before matching colors for each individual parameter, while the other kits had a single wait time for all parameters. One kit (M4) provided a plastic test tube and guidance on sample volume.

Table 4 | Characteristics of the evaluated OTS test kits (at the time of data collection in 2018)

Test kit brand	Measures					Costs			Guidance	
	Fe	Cu	Mn	FI	Add	\$/cont	#/cont	\$/test	V	Flush
M1	Y	Y	N	Y	^a	18.99	100	0.19	N	N
M2	Y	Y	N	Y	^a	19.99	125	0.16	N	N
M3	Y	Y	N	Y	^{a, b}	19.95	150	0.13	N	N
M4	Y	Y	N	Y	^{a, b}	18.99	100	0.19	Y	N
SA1	Y	N	N	N	–	25.25	25	1.01	Y	N
SB2	Y	N	N	N	–	18.88	50	0.38	N	N
SC3	Y	N	N	N	–	19.99	25	0.80	N	N
SD4	Y	N	N	N	–	26.54	50	0.53	Y	N
SA5	N	Y	N	N	–	12.94	25	0.52	N	N
SC6	N	Y	N	N	–	13.29	25	0.53	Y	N
SC7	N	N	Y	N	–	28.31	24	1.18	Y	N

Note. Add (additional measures): ^a multiparameter kit that measures free chlorine residual, lead, total alkalinity, and total hardness, nitrate, nitrite, pH. ^b Multiparameter kit that measures bromine, carbonate, cyanuric acid, and total chlorine. V = sample volume; flush = instructions about flushing (letting water run through pipes for a specified amount of time) and/or when to take a sample were included; cont = sample container was included in the kit.

Three of the iron-only test kits (SA1, SB2, and SC3) and all of the copper-only kits involved dipping a strip in the water sample and matching a color on the strip to a color chart provided on the test kit bottle after a specified amount of time. The iron-only SD4 had users match the color of the liquid in the test tube with the colors on a color chart. Three of the iron-only kits provided a sample vial (the SA1 kit provided a 100 mL bottle, the SC3 a 5 mL bottle, and the SD4 a 5 mL plastic test tube), and two of these kits instructed users to add a reducing agent to the sample (SD4 and SA1). The copper-only SA5 did not instruct users on a sample volume, while the SC6 instructed users to test 200 mL of water (although no sample vial was provided). Neither copper-only kit provided guidance on flushing or the time of day to take a sample. The manganese kit (SC7) involved dipping test strips and matching them to a color chart and providing a 5 mL plastic sample tube.

The kits offered varying levels of information on how to interpret their results (Table S2). The M2 and M3 had boxes around the ‘ideal range’ or ‘target level’, respectively, for each parameter. The M4 had the ‘optimal level’ for each parameter written in green versus written in black for all other levels, while the M1 included words such as ‘OK’, ‘high’, and ‘low’ above the concentration levels on the color chart. The provided information was not consistent: for instance, for iron, the M1 instructions informed users that 0–0.3 ppm was ‘OK’ and 0.5–5 ppm was ‘high’, while the remaining multiparameter test kits informed users that 0 ppm was ‘ideal’. For copper, the M1 defined 1–2 ppm as ‘high’ but the M3 has 1 mg/L as a ‘target’.

Trial with untrained users

Accuracy

The untrained users’ results yielded similar results by individual and group; however, there were differences in the results obtained by measurements made by untrained users compared to lab measurements (Figures 5 and S2). In the DI water matrix under laboratory concentrations (Figure 1), the M3 lab kit consistently underestimated concentrations, while under untrained user conditions, it overestimated concentrations (Figures 5 and S2). Interestingly, in Amherst tap water, the M3 kit measured 0 mg/L in the laboratory-obtained data (Figure 1), but the untrained users did detect iron at a higher concentration of 12.3 mg/L (although it was underestimated as 5 mg/L).

While in the DI water matrix under lab conditions, the SA1 performed consistently well (Figure 2), but when used by untrained users, the measurements underestimated concentrations (Figures 5 and S2). In the tap water matrix, the SA1 laboratory-obtained and untrained users’ data were similar (Figures 1, 5, and S2). Similarly, in the DI water, the SC3 led to overestimates of concentrations under lab conditions but underestimates when measured by untrained users (Figures 1 and 5).

Out of three kits trialed, users rated the only multiparameter kit (M3) the highest for clear instructions (Table 5). The SA1 was rated as the easiest to read results and most confidence in results (Table 5) (notably, this kit was the only one requiring the

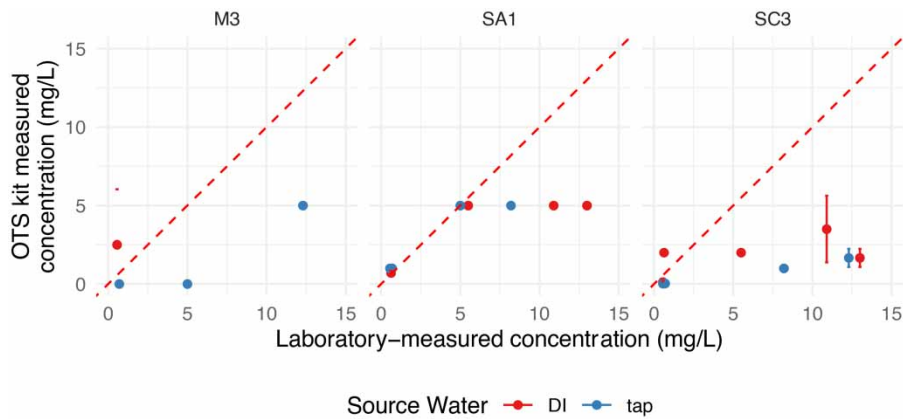


Figure 5 | Iron concentrations (0–12 mg/L) measured by untrained users (y-axis, OTS kits measured) in the tap water matrix compared to laboratory-obtained values (FerroVer Iron Reagent) on the x-axis. The diagonal lines represent perfect matches. Values >12 mg/L are not shown; these included two laboratory-measured values of 96 mg/L (which citizen scientists measured as 2 mg/L with SC3 and SA1) and six samples of 118 mg/L, which were measured by participants as 2 mg/L (SC3), 5 mg/L (SA1, three samples), and 50 mg/L (M3, two samples).

Table 5 | Average results from surveys

	The instructions were clear	The results were easy to read	I have confidence in the results	I know what these results mean for the safety of the water
SC3	4.65	3.20	3.25	3.90
SA1	4.58	4.46	4.21	3.17
M3	5.00	3.67	3.22	3.44

Note. Each column represents a statement and each row is a test kit brand. SC3 $n = 20$. SA1 $n = 24$. M3 $n = 9$. The highest average response for each question is highlighted in green. Untrained users were asked to rate each statement on a scale of 1–5. 1 = strongly disagree, 3 = neutral, and 5 = strongly agree.

addition of any reagent), while the SC3 kit was rated highest for understanding what the results meant for water safety (this kit identified the EPA standard on the bottle). Overall, the users found the kit instructions clear, but their ability to give meaningful and interpretable results was lacking (Table 5).

DISCUSSION

Our results showed highly variable performance in the test kits' ability to detect potential contaminants in the water. In the measurements, results showed that results were largely consistent; therefore, test kit results were generally highly precise but not always accurate. Accuracy was often affected by the water matrix: in general, there were few noticeable differences in performance between tests performed in tap water compared to river water, while differences were frequently observed when comparing results in DI water as compared to the tap or river water. Notably, no one brand or kit performed consistently well across multiple water quality parameters; the same brand in some cases performed well for one parameter and poorly for another, or the same multiparameter kit was able to adequately measure one parameter but not another. In general, the test kits that measured only single parameters performed better than the multiparameter kits when measuring iron or copper. Two of the test kits measuring only iron that performed well included a reducing agent, which would reduce any ferrous iron to ferric iron. It is important to consider the intervals at which each kit measures: a concentration measured by a test kit that was markedly different from the laboratory-obtained concentration may have been limited not by the intensity of color it produced but by the design of the OTS kit as having only the ability to measure in measurement bins (e.g. 0–5 mg/L rather than the ability to differentiate 1 vs. 4 mg/L).

Our results are consistent with those found by studies focused on the measurement of other contaminants, which have also identified challenges that home users had or would have with obtaining accurate results, such as for nitrate or lead, showing

that false negatives or user-measured results as lower were common (Nielsen *et al.* 2008; Kriss *et al.* 2021). The observed differences in performance are likely due to test kit chemistry and water matrix interference. While the test strip kit chemistries are not known (information was not disclosed by manufacturers), our results demonstrated that the iron measurement test kits involving reducing agents performed better as they could then likely detect both ferrous and ferric iron (similar challenges with metal speciation detection via home test kits have been previously observed with the measurement of lead (Kriss *et al.* 2021)). Multi-parameter test kits often performed worse than single-parameter kits, potentially as they tried to optimize for multiple analytes at once, many of which may need different pH for optimum performance. Also, as noted in the literature on colorimetric methods, interference from the water matrices likely affected kit performance; in our experiments, performance in tap and river water was poorer than in DI water, suggesting interference from organic matter and other ions likely occurred. However, notably, our laboratory-determined values also relied on colorimetric methods. Additionally, we selected sulfate salts as the source compounds for iron, copper, and manganese; however, to facilitate future standardization for evaluating commercial test kit performance, future work could compare kit performance with the addition of different anions to the ability to mimic those found in drinking water, particularly where water matrices may influence results, as well as better understand speciation and its effect on results.

Some of the multiparameter kits outperformed the parameter-only kits if the goal was a binary outcome of above or below the SMCL; while the copper-only kits may be more accurate, the sensitivity analysis revealed that some of the multiparameter kits outperformed the copper-only kits when assessing their ability to compare to the SMCL. Therefore, the selection of test kits should consider its application: for a water consumer, the ability of a test kit to meaningfully provide information about a water sample's compliance with an MCL may be more important than accuracy or the ability to detect >0 mg/L.

This study had limitations that also suggest future directions. Our 'true value' test was with another colorimetric method; while a more accurate method such as ICP-MS would have yielded more accurate results for our true value, we selected the single-parameter colorimetric method as our base of comparison as many water utilities and scientific studies are conducted with this method, and this is what the test strips are trying to emulate. For example, the kit testing manganese was a direct modification of the Hach PAN LR spec method, with the steps and reagents closely matching each other. Future work could also compare to results from an ICP-MS. For these experiments, tap and river water were sometimes collected on different days due to logistical constraints, which would result in different background concentrations of parameters of interest or constituents that could potentially interfere with results; however, these sources were not expected to vary by much over the few days in which water was collected, and we measured the background concentration of the parameter of interest on each day water was collected. Future studies should also incorporate testing in different waters, particularly groundwater, as those are the sources that many US homeowners use, as well as with varying pH and temperature, which would be expected to significantly affect results. Future work should investigate the constituents in source waters that could be causing interference with the test kits, such as through the use of synthetic water and testing potential interference (e.g. organic matter, alkalinity, hardness, and pH) and conducting more detailed analysis of water matrix chemistry. While consistent lab personnel were used in this study, factors such as time of day, lighting conditions, colorblindness, and eyesight may affect results, although the kit instructions did not mention these factors. A procedure for consistently reading color charts could help eliminate this large source of error, such as other efforts to use a mobile phone to read the results of colorimetric methods. Future work should also focus on varying some of the key method procedures to reproduce mistakes likely made by potential users, such as waiting more or less time than instructed to dip the strip in the water or matching the color. Finally, we tested only a very small fraction of the test kits that are available on the market, and the selection was nonrandom; we sought to replicate the procedure a home user may use when selecting test kits and acknowledged that manufacturer and kit variability changes daily. Future studies could focus on more holistically evaluating kits on the market and other procedures for selecting test kits, such as by kit chemistry or aligned result bins, as well as evaluating the many other parameters that users may want to measure.

Overall, these test kits are being sold and used by consumers across the United States. However, our results suggest that there are concerns about their accuracy and usefulness and that whether the results can be relied on depends on which parameter is being measured in which water with which kit and for which purpose. While many of these studies have focused on the assessment of commercially available test kits, recent efforts have highlighted opportunities to co-design test kits with users such as middle and high school teachers (Haynes *et al.* 2019). Previous studies focused on preparing and validating field test kits, particularly for use in low- and middle-income countries to measure chlorine or fecal indicator bacteria, have shown that, with method development and evaluation of user experiences, such kits can be reliable and accurate;

however, many of these are still focused on use by those with some training in water quality rather than home users (Bain *et al.* 2012; Khush *et al.* 2013; Murray & Lantagne 2015; Bain *et al.* 2021). The results of our study suggest a need to develop standardized protocols for evaluating the appropriateness and validity of these home test kits given different goals and water sources. There are significant gaps in our knowledge of the performance of these test kits and how to engage with potential users in selecting appropriate parameters to measure and in interpreting the results into actionable steps. Considering both the reliability and accuracy of the results, as well as how potential users may interact with the kits, is important for home users, community groups, or citizen science initiatives that may be relying on these results.

ACKNOWLEDGEMENTS

We thank Emily Bonaccorso for her assistance in the laboratory and feedback from David Reckhow.

FUNDING

We acknowledge funding from the University of Massachusetts Amherst Civil and Environmental Engineering.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- ANSI/AWWA (2011) *ANSI/AWWA B701-11 Sodium Fluoride*. Denver, CO: American Water Works Association.
- Aukema, K. G. & Wackett, L. P. (2019) *Inexpensive microbial dipstick diagnostic for nitrate in water*, *Environmental Science: Water Research & Technology*, **5**, 406–416. <https://doi.org/10.1039/C8EW00834E>.
- Bain, R., Bartram, J., Elliott, M., Matthews, R., McMahan, L., Tung, R., Chuang, P. & Gundry, S. (2012) *A summary catalogue of microbial drinking water tests for low and medium resource settings*, *International Journal of Environmental Research and Public Health*, **9**, 1609–1625. <https://doi.org/10.3390/ijerph9051609>.
- Bain, R., Johnston, R., Khan, S., Hancioglu, A. & Slaymaker, T. (2021) *Monitoring drinking water quality in nationally representative household surveys in low- and middle-income countries: cross-sectional analysis of 27 multiple indicator cluster surveys 2014–2020*, *Environmental Health Perspectives*, **129**, 097010. <https://doi.org/10.1289/EHP8459>.
- Flanagan, S. V., Marvinney, R. G. & Zheng, Y. (2015) *Influences on domestic well water testing behavior in a Central Maine area with frequent groundwater arsenic occurrence*, *Science of The Total Environment*, **505**, 1274–1281. <https://doi.org/10.1016/j.scitotenv.2014.05.017>.
- George, C. M., Zheng, Y., Graziano, J. H., Rasul, S. B., Hossain, Z., Mey, J. L. & van Geen, A. (2012) *Evaluation of an arsenic test kit for rapid well screening in Bangladesh*, *Environmental Science & Technology*, **46**, 11213–11219. <https://doi.org/10.1021/es300253p>.
- Haynes, E. N., Hilbert, T. J., Roberts, R., Quiroigico, J., Shepler, R., Beckner, G., Veevers, J., Burkle, J. & Jandarov, R. (2019) *Public participation in air sampling and water quality test kit development to enable citizen science*, *Progress in Community Health Partnerships: Research, Education, and Action*, **13**, 141–151.
- Khush, R. S., Arnold, B. F., Srikanth, P., Sudharsanam, S., Ramaswamy, P., Durairaj, N., London, A. G., Ramaprabha, P., Rajkumar, P., Balakrishnan, K. & Colford, J. M. (2013) *H₂S as an indicator of water supply vulnerability and health risk in low-resource settings: a prospective cohort study*, *American Journal of Tropical Medicine and Hygiene*, **89**, 251–259. <https://doi.org/10.4269/ajtmh.13-0067>.
- Kriss, R., Pieper, K. J., Parks, J. & Edwards, M. A. (2021) *Challenges of detecting lead in drinking water using at-home test kits*, *Environmental Science & Technology*, **55**, 1964–1972. <https://doi.org/10.1021/acs.est.0c07614>.
- Murray, A. & Lantagne, D. (2015) *Accuracy, precision, usability, and cost of free chlorine residual testing methods*, *Journal of Water and Health*, **13**, 79–90. <https://doi.org/10.2166/wh.2014.195>.
- Murray, A., Hall, A., Weaver, J. & Kremer, F. (2021) *Methods for estimating locations of housing units served by private domestic wells in the United States applied to 2010*, *JAWRA Journal of the American Water Resources Association*, **57**, 828–843. <https://doi.org/10.1111/1752-1688.12937>.
- National Research Council (1997) *Safe Water From Every Tap: Improving Water Service to Small Communities*, *Committee on Small Water Supply Systems*. Washington, DC: National Academies Press. <https://doi.org/10.17226/5291>.
- Nielsen, S. S., Mueller, B. A. & Kuehn, C. M. (2008) *An evaluation of semi-quantitative test strips for the measurement of nitrate in drinking water in epidemiologic studies*, *Journal of Exposure Science & Environmental Epidemiology*, **18**, 142–148. <https://doi.org/10.1038/sj.jes.7500563>.

- Patel, A. I. & Schmidt, L. A. (2017) Water access in the United States: health disparities abound and solutions are urgently needed, *American Journal of Public Health*, **107**, 1354–1356. <https://doi.org/10.2105/AJPH.2017.303972>.
- Pierce, G. & Gonzalez, S. (2017) Mistrust at the tap? factors contributing to public drinking water (mis)perception across US households, *Water Policy*, **19**, 1–12. <https://doi.org/10.2166/wp.2016.143>.
- Pierce, G., Gonzalez, S. R., Roquemore, P. & Ferdman, R. (2019) Sources of and solutions to mistrust of tap water originating between treatment and the tap: lessons from Los Angeles County, *Science of The Total Environment*, **694**, 133646. <https://doi.org/10.1016/j.scitotenv.2019.133646>.
- Powers, M., Yracheta, J., Harvey, D., O’Leary, M., Best, L. G., Black Bear, A., MacDonald, L., Susan, J., Hasan, K., Thomas, E., Morgan, C., Olmedo, P., Chen, R., Rule, A., Schwab, K., Navas-Acien, A. & George, C. M. (2019) Arsenic in groundwater in private wells in rural North Dakota and South Dakota: water quality assessment for an intervention trial, *Environmental Research*, **168**, 41–47. <https://doi.org/10.1016/j.envres.2018.09.016>.
- R Core Team (2024) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Reddy, R. R., Rodriguez, G. D., Webster, T. M., Abedin, M. J., Karim, M. R., Raskin, L. & Hayes, K. F. (2020) Evaluation of arsenic field test kits for drinking water: recommendations for improvement and implications for arsenic affected regions such as Bangladesh, *Water Research*, **170**, 115325. <https://doi.org/10.1016/j.watres.2019.115325>.
- Srivastava, S. & Flora, S. J. S. (2020) Fluoride in drinking water and skeletal fluorosis: a review of the global impact, *Current Environmental Health Reports*, **7**, 140–146. <https://doi.org/10.1007/s40572-020-00270-9>.
- Tang, M., Nystrom, V., Pieper, K., Parks, J., Little, B., Williams, R., Esqueda, T. & Edwards, M. (2018) The relationship between discolored water from corrosion of old iron pipe and source water conditions, *Environmental Engineering Science*, **35**, 943–952. <https://doi.org/10.1089/ees.2017.0435>.
- US EPA (2015) SW-846 Test Method 9214: Potentiometric Determination of Fluoride in Aqueous Samples with Ion-Selective Electrode (Other Policies and Guidance). Washington, DC: Environmental Protection Agency.
- US EPA (2015a) *Information About Public Water Systems [WWW Document]*. US EPA. Available at: <https://www.epa.gov/dwreginfo/information-about-public-water-systems> (Accessed: 15 November 2020).
- US EPA (2015b) *National Primary Drinking Water Regulations [WWW Document]*. National Primary Drinking Water Regulations. Available at: <https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations> (Accessed: 13 June 2024).
- Vidmar, A. M., Wells, E. C., Zheng, M., Awad, N., Combs, S. & Diaz, D. (2023) ‘That’s what we call ‘aesthetics,’ not a public health issue’: the social construction of tap water mistrust in an underbounded community, *Human Organization*, **82**, 342–353. <https://doi.org/10.17730/1938-3525-82.4.342>.

First received 2 August 2024; accepted in revised form 28 February 2025. Available online 7 March 2025