Short study on the VER generation of Vestergaard Frandsen's LifeStraw Carbon For Water programme with special regard to usage time of water filters

Matthias Krey

Final report

Submitted to:

atmosfair

Freiburg, Germany, 27.09.2017

Perspectives
Climate Group GmbH
Hugstetter Str. 7
79106 Freiburg, Germany
info@perspectives.cc
www.perspectives.cc

perspectives climate group

Contents

1 Background

In June 2011, the Vestergaard Frandsen's LifeStraw Carbon For Water programme in Kenya (GS886) registered under the Gold Standard (in the following: "Vestergaard GS886") to generate Gold Standard Verified Emission Reductions (VERs). As part of the programme 877,505 household water filters, so called LifeStraw filters (LSF) have been distributed free of charge to households, who would otherwise boil water to treat water for consumption on stoves using high carbon emission fuels. By the end of the distribution campaign, 91% of all households in the Western Province of Kenya had received a LSF. Since 2010, the programme is heavily criticized by international aid workers, NGOs and the media (Starr 2011, Heinemann 2013, Atmosfair et al. 2013). atmosfair gGmbH (in the following: "atmosfair"), a Gold Standard NGO supporter, has repeatedly expressed its concerns regarding the environmental integrity of the programme and the large amount of VERs issued. In June 2016, atmosfair requested the Gold Standard to launch a formal grievance procedure against the programme asking for a detailed investigation in the project's environmental integrity. The Gold Standard explained that it had previously investigated the issue in 2013 with support from Berkeley Air Monitoring Group and that the outcome of this investigation was a new guideline for sampling of LSF projects and adjustments to the amounts of VERs issued for the Vestergaard GS886. Gold Standard closed the grievance. But when in September 2016 the Stanford University study "Climate and Health Co-Benefits in Low-Income Countries: A Case Study of Carbon Financed Water Filters in Kenya and a Call for Independent Monitoring" found considerably lower usage rates among participants of Vestergaard GS886 than reported by Vestergaard Frandsen, the Gold Standard launched a second grievance procedure to investigate whether the Gold Standard had over-issued VERs to Vestergaard GS886. This process is currently ongoing. atmosfair has asked Perspectives Climate Group (in the following: "PCG") to validate its criticism of Vestergaard GS886, particularly relating to the conclusions drawn from the studies of the Berkeley Air Monitoring Group and Stanford University. As atmosfair has also estimated the amount of VER overissuance to Vestergaard GS886, it also requested PCG to validate the approach, methodology used and results obtained by atmosfair for this calculation.

2 Criticism of Vestergaard GS886 VER issuances

One of the main arguments for critiques to shun Vestergaard GS886 is the concept of so called suppressed demand. It allows the programme to generate emissions reductions based on an assumed volume of water treatment (that is currently suppressed due to poverty or lack of water treatment infrastructure), that is considerably higher than the current water filter use of programme participants. It needs to be noted that the principle of suppressed demand is widely used in market mechanisms for generation of emission credits in developing countries, e.g. under the Clean Development Mechanism (CDM).

The second major criticism is, in the eyes of the contenders, the unrealistically high value applied for the utilisation rate of the LSFs after distribution (that was periodically determined through surveying programme participants), as the amount of VERs that the programme generates strongly depends on this parameter. As a response to the severe criticism the Gold Standard contracted the Berkeley Air Monitoring Group in 2013 to independently investigate the issue (Gold Standard 2016). The findings eventually led to Gold Standard mandating adjustments to the usage rates determined during the second monitoring period of the Vestergaard GS886. The amount of VERs issued compared to the initial verification report for this period were reduced by 449,539, i.e. 21%. As a further reaction to this challenge, Gold Standard in January 2014 released the "Guidelines for carrying out usage surveys for projects implementing household water filtration technologies" that consider best-practice sampling methods as described in the Berkeley Air Monitoring Group report (Gold Standard 2014). According to the Gold Standard the guidelines have been taken into account for the third monitoring period of the Vestergaard GS886 (Gold Standard 2016).

The table below shows the amounts of VERs issued to the programme for the corresponding monitoring periods up to date as well as the underlying values for the LSF usage rates.

Milestone	Date	Monitoring Period	GS VERs issued	Usage rate
1st verification report	February 2012	01.06.2011 - 30.11.11	1,355,185	91.13%
2nd verification report	December 2013	01.12.11 - 31.10.12	1,701,562 (2,151,101)*	74.98% (92.76%)*
3rd verification report	June 2014	01.11.2012 - 31.01.2014	1,419,458	80.46%

* Numbers in brackets had been initially verified before Gold Standard mandated changes (and were eventually not issued)

Table 1: GS VER issuances Vestergaard GS886

The Stanford study of 2016 reports 19% usage of the LSFs 2-3 years after filter distribution (among households with pregnant women that have received LSFs through the programme) compared to the 80% usage reported by the Vestergaard GS886 around 2.7 years after filter distribution (Pickering at al. 2016). One factor that could have caused this significant discrepancy in the results is the way that

the surveys have been set-up and executed and how this might have impacted the results obtained. Therefore, the surveys are analyzed in detail in the following section.

3 Analysis of the surveys

3.1 Discussion on minimum sample size for sampling usage rate

Sampling is applied for the determination of the value of specific parameters in case the parameter has to be determined based on a large number of elements. By applying a sampling approach, fewer elements have to be evaluated. The parameters of interest can be divided into two categories as shown in the figure below.

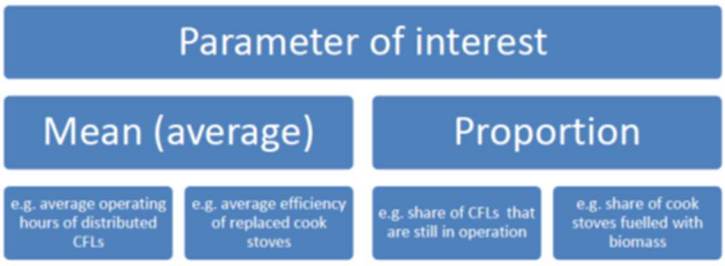

Figure 1: Parameters of Interest in the Sampling Approach

Source: Feige et al. (2012)

Proportions are calculated based on observations where an attribute of a specific parameter is either true or false. The usage rate in the Vestergaard GS886 (U_y) is such a proportion, as it describes the percentage of households that use the LSFs distributed in the programme.

Sampling is therefore a valid approach for determining the usage rate in Vestergaard GS886. But it has to be taken into account that the results obtained for the usage rate in a sample of households that have received the LSFs is never going to be identical to the real usage rate of the total population of all 877,505 households that participate in the programme. It is only possible to get as close as possible to the real usage rate by increasing the size of the sample. In fact, statistics provide a particular formula to calculate the minimum sample size for representative sample results (Feige et al. 2012). When applying this formula and best-practice standard parameters, the minimum sample size for determining the usage rate in Vestergaard GS886 should be between 43 households (when assuming that 90% usage of LSFs will occur) and 3.444 households (when assuming that 10% usage of LSFs will occur).

The underlying GS methodology requires a survey with the following minimum sample sizes to determine the usage rate (Gold Standard 2010):

- Population size < 300: Minimum sample size 30
- Population size 300 to 1000: Minimum sample size 10% of group size
- Population size > 1000: Minimum sample size 100

This means that the methodology requires a minimum sample size of 100 households for Vestergaard GS886 to determine the usage rate. This corresponds to an expectation that 80% usage of LSFs will occur in the programme, as the above mentioned formula requires a minimum sample size of 97 households when assuming that 80% of households will use the LSFs they have received. This is in line with the results of a pilot sampling campaign reported in the PDD that showed 83% of usage of LSFs after distribution.

3.2 Criteria for analysing surveys

For systematically analysing the quality of the surveys and for making them comparable with each other, an information sheet that includes the following criteria is used for each survey.

Criteria	**For determining…**
Characteristics of the sample population	…how representative is the sample of the total population that received LSF filters under the programme?
Sample size (households)	…if the sample size is sufficiently high to yield representative results?
Sample size (households)	…if the sample size is sufficiently high to yield representative results?
Approach	…how suitable is the approach of the survey for determining behavioral practices among the households surveyed (e.g. self-reporting less suitable than objective observations)? …if random-sampling been used to ensure that the sample population is representative of the total population?
Methodology	…,e.g. …if the survey questions are phrased in such a way that they provide room for multiple answers (open questions) and are not leading questions …if there is a logical connection between the questions, the type of answers that are expected and the parameter that is supposed to be sampled

	…if the setting of the interview can avoid courtesy bias (tendency of respondents to give answers that they think the interviewer wants to hear) to highest degree possible?
Surveyor	…if the surveyor is free of a conflict of interest to ensure there is no incentive to misreport?
Quality control	…if spot-checks have been carried out to validate sample results?

Table 2: Criteria for determining the quality of surveys and assuring comparability

Source: Own illustration

3.3 Vestergaard GS886 surveys

Vestergaard GS886 – 1st verification - VF-MR1-Survey

Usage rate	91.13%
Survey period	July and August 2011
Characteristics of the sample population	Households that have obtained an LSF under the programme (is approximately 80% women between 15 and 64) Households from across all 32 districts in the project region
Sample size (households)	19,430
Approach	Most likely not a random sample Self-reporting
Methodology	Surveys were conducted one week following education campaigns across several districts. During the survey, the usage was confirmed by asking households to demonstrate how they use the Lifestraw unit. Correct usage is indicative of routine use. There are a number of questions in the survey that would indicate whether the unit is used by the household: • Does the person understand that they should use LifeStraw filtered water for drinking every time and every day? • When is the last time you used LifeStraw to filter water?

	• How much water do you filter using LSF in one day? (explaining by using the size of Jerrycan) • How often do you filter water? How many times do you fill the container in a day? • How many liters of filtered water does your family use for drinking, washing fruits and vegetables and hand washing every day? Participants that 1) successfuly demonstrated using their Lifestraw unit and 2) who reported using their LifeStraw Family at least 2 times per week were counted as users
Surveyor	Vestergaard Frandsen
Quality control	Survey conducted by EXP Agency of 101 households to quality-check data (sample from a random sample of 382 households selected from the households visited during the VF survey with representation from each District) finding higher U than VF-MR1 (98.02%) ERM CVS (DOE) sampled 18 VF-MR1-Surveyed households and confirmed the answers provided to those questions related to usage finding 100% usage

Table 3: Assessment of quality criteria for the Vestergaard GS886 first verification report

Source: Own illustration based on Vestergaard Frandsen (2012)

The following observations which negatively influence the quality of this survey have to be made:

- The sampling was very likely not a random sample (as otherwise this would have been stated in the MR or VR, as random sampling is required in the underlying GS methodology)
- Strong reliance on self-reported information
- It is not adequate to automatically assume that a demonstration of correct use by the person means that he routinely uses the LSF
- No definition of a “successful demonstration”. This leaves a lot of room for interpretation of when criteria 1) for a user is fulfilled
- No clear link between questions and how the answers to those questions are used to determine criteria 2) (“who reported using their LifeStraw Family at least 2 times per week”) for a user

- The survey was undertaken one week after the end of the education campaign which means that it is reasonable to assume that a large number of households would well recall how to use the LSF
- The data was collected by Vestergaard Frandsen (probably the employees that also conducted the education campaign that ended a week before the survey). This could have created a large incentive for overreporting as the results also assess the efficacy of the education campaign
- The 101 cross-check households have been surveyed by EXP Agency, which is a marketing firm specialized on user activation that was also hired by Vestergaard Frandsen for conduction awareness-raising campaigns and stakeholder consultations during the project implementation. Therefore, EXP Agency could have had an incentive to overreport.
- The sample size of 18 households to conduct spot-checks is not sufficiency to yield representative results

Vestergaard GS886 – 2nd verification - VF-MR2-Survey

	Initial MR February 2013	Final MR December 2013
Usage rate	92.76%	74.98%
Survey period	Survey 1: April 11 - May 24, 2012 Survey 2: October 15 - 31, 2012	Usage rate not determined through surveys/sampling!
Characteristics of the sample population	See VF-MR1-Survey	-
Sample size	**20,220** (13,308 from Survey 1 and 6,912 from Survey 2)	-
Approach	See VF-MR1-Survey	-
Methodology	See VF-MR1-Survey	-
Surveyor	See VF-MR1-Survey	-

Quality control	Survey conducted by EXP Agency of 252 households ERM CVS (DOE) sampled 35 VF-MR2-Surveyed households and confirmed the answers provided to those questions related to usage finding 100% usage	-

Table 4: Assessment of quality criteria for the second verification report

Source: Own illustration based on Vestergaard Frandsen (2013)

As can be seen above, the VF-MR2-Survey had very similar flaws as the VF-MR1-Survey (with the only difference that the size of the sample size and the quality control groups were (slightly) higher).

In 2013, Gold Standard contracted the Berkeley Air Monitoring Group to independently investigate potential flaws in the approach for determination of the usage rate based on the approach used for the 1st and 2nd monitoring period. The Berkeley findings were available in November 2013 (Graham et al. 2013). The main criticism of the survey approach taken by Vestergaard Frandsen was the strong reliance on self-reported data, generally agreed upon by experts as poor proxy measures of use. The report estimated the appropriate usage rate based on water, sanitation, and hygiene (WASH) sector expert interviews about the quality on the questions asked by Vestergaard Frandsen and the responses provided. In doing so, the study makes a more stringent definition of non-users using the following criteria:

- reported not having LSF filtered water in their safe storage container
- needed a replacement part
- could not demonstrate use or backwash
- reported never backwashing/cleaning
- did not report blue tap as for safe water

Based on these criteria the study recommends a usage rate of 74.98%

Table 6: Usage rates by question for Corrected Processing[4]

Corrected Processing Q[April-May]/[October]	Total (minus outliers)	Reported Rate	Total Non-Users	Additional Non-Users	Total Users	Usage Rate
Original	20220	9.6%	1940	0	18280	90.41%
Q12/_ No LSF water in safe storage container	20220	10.2%	3358	1418	16862	83.39%
Q13/24 Cannot demonstrate use	20220	6.9%	2097	157	18123	89.63%
Q16/25 Cannot demonstrate backwash	20220	14.9%	3458	1518	16762	82.90%
Q13&16/24&25 Cannot demonstrate at least one	20220	15.8%	3615	1675	16605	82.12%
Q17/26 Report never backwash/clean	20220	3.2%	2019	79	18201	90.01%
Q18/_ Don't report Blue tap as safe water	20220	0.3%	1965	25	18255	90.28%
Q5/13 Filter not hanging correctly	20220	5.0%	2530	590	17690	87.49%
Q_/14 Had to unblock pre-filter	20220	3.7%	2507	567	17713	87.60%
Q_/15 Had to unblock cartridge	20220	3.5%	2457	517	17763	87.85%
Q_/14&15 Had to unblock pre-filter or cartridge	20220	5.2%	2725	785	17495	86.52%
Q_/16 Need filter part replacement	20220	2.1%	2198	258	18022	89.13%
Q6/18 No safe storage container	20220	5.2%	2616	676	17604	87.06%

Table 7: Usage rates of categories

Corrected Processing	Total (minus outliers)	Reported Rate[5]	Total Non-Users	Additional Non-Users	Total Users	Usage Rate
Original	20220	-	1951	0	18269	90.40%
VF Reported WHO Toolkit[6]	20220	11.0%	2790	850	17430	86.20%
Additional WHO Toolkit[7]	20220	25.8%	5278	3338	14942	73.90%
All combined[8]	20220	32.4%	6391	4451	13829	68.39%
Recommended[9]	20220	21.9%	5060	3120	15160	74.98%

Table 5: Usage rates by question for Corrected Processing and Usage rates of Categories

Source: Graham et al. (2013)

However, this recommended rate does not take into account that 5% of the households did not have the LSF hanging properly. As the new guidelines for determining usage rates released in 2014 included this criterion in the assessment of the usage rate (Gold Standard 2014), it should also be taken into account for the monitoring period 2 usage rate. The resulting and more conservative usage rate would be 72.05%.

One WHO expert suggested the following questions to assess usage of the LSF, where the respondent would be classified as a user if he/she responded in the affirmative to all the questions/observations listed. The resulting usage rate for this survey methodology cannot be determined for monitoring period 1 and 2, because the surveys cannot be repeated again. We will however come back to this methodology later when analyzing the VF-MR3-Survey.

Survey questions that would provide rigorous, yet conservative, estimates of LifeStraw Family filter use.

Self-Reported Usage Measures			
Q 1	Do you do anything to make your water safe for drinking?	Yes No Don't know Refuse to answer	 Skip to Q Skip to Q Skip to Q
Q2	What do you do to make your water safe for drinking?	Boil Use LSF filter Use chlorine Other:_____	Skip to Q Skip to Q Skip to Q
Q3	Can you provide me a cup of water that you would prepare for your child?	Yes No Don't know Refuse to answer	 Skip to Q Skip to Q Skip to Q
Q4	Did you do anything to make this water safe for drinking?	Yes No Don't know Refuse to answer	 Skip to Q Skip to Q Skip to Q
Q5	What did you do to make this cup of water safe for drinking?	Boil Use LSF filter Use chlorine Other:_____ Don't know Refuse to answer	Skip to Q Skip to Q Skip to Q Skip to Q Skip to Q
Q6	Is your LifeStraw Family Filter working now?	Yes No Don't know Refuse to answer	 Skip to Q Skip to Q Skip to Q
Observational Usage Measures			
Q7	May I observe your LifeStraw?	Yes No Not observable	 Skip to Q Skip to Q
Q8	Is the LSF hanging correctly, with ropes positioned to allow the pre-filter to come out?	Yes No	 Skip to Q
Q9	Is the filter wet or moist?	Yes No	 Skip to Q

Figure 2: Survey questions that would provide rigorous yet conservative estimates of LSF filter use

Source: Graham et al. (2013)

Finally, the following point is noteworthy regarding the MR and VR of monitoring period 2.

Even though the final VR report states that…

"*after recommendations by the external experts [comment by the author: this refers to the Berkeley study] on the assessment methods were provided, the following survey's questions were used:*

- *What do you use to make your water safe now? (without reading out the choices)*
- *Does the safe storage container have LifeStraw filtered water in it?*
- *Can the person demonstrate how to filter water using the LifeStraw*
- *correctly?*
- *Can the person you are interviewing demonstrate how to backwash*
- *the LifeStraw correctly?*
- *How often do you backwash and clean the pre-filter? (without suggesting choices)*
- *Which tap is used for safe water?*

- *How many liters of filtered water does your family use for drinking, washing fruits and vegetables and hand washing each day? (based on amount of the jerry can used)*
- *How often do you use the LifeStraw filter? (without suggesting choices) …*"

Source: ERMCVS (2014a)

…no new survey has been conducted using the guidance as suggested by the Berkeley study, as the 2nd (and latest) survey mentioned in the MR/VR ended October 2012. This is further supported by the fact that the Gold Standard decided to issue a limited amount of VERs "pending further investigation (based on 55% conservative usage rate)" on 16th May 2013 (Gold Standard 2016). Gold Standard later justified this with "delivery commitments" of VF (Gold Standard 2016). Gold Standard did not provide any justification how the 55% usage rate has been determined and why the Gold Standard thought it was a "conservative" value.

As the final usage rate of 74.98% for MP2 has not been determined through a new survey, it seems therefore most realistic to assume that the Gold Standard simply asked Vestergaard Fransen to accept the "best estimate" provided in the Berkeley study, even though the final MR report tries to create the impression that the new guidance by the Berkeley study was taken into account in the surveys.

Vestergaard GS886 – 3rd verification - VF-MR3-Survey

Usage rate	80.46%
Survey period	December 7, 2013 - January 31, 2014
Characteristics of the sample population	Households that have obtained a LSF under the programme (is approximately 80% women between 15 and 64)
Sample size	16,313
Approach	Self-reporting and observations
Methodology	A list of question for the interviews is not provided in the MR/VR. Instead a table explaining the process steps for determining the usage rate is included in the VR report. The steps that relate to the actual way the interviews/observations were conducted are the following: • Removed households that did not report "LifeStraw" as the method used to make their water safe ("What do you do to make your water safe now?") • Removed any households where the filter was not hanging properly at the time of visit, and there was not a legitimate reason why it was not hanging • Removed households that did not have filtered water in the storage container within the last 24 hours • Removed households that filtered less frequently than once every two weeks. • Removed households that could not demonstrate how to filter

	• Removed households that could not demonstrate how to backwash • Removed households with filters that are not functioning • Removed households that do not clean the pre-filter or backwash at least once every two weeks The complete table is provided in the Annex (include). The third education campaign under the project took place in October-November 2012, overlapping with this monitoring period.
Surveyor	Volunteers recruited by Vestergaard Frandsen which received a nominal fee as payment and Sub-county coordinators employed by Vestergaard Frandsen
Quality control	Survey conducted in by EXP Agency of 257 households during January, 2014 reported "Vestergaard Frandsen: Lifestraw Family Monitoring Survey Report Western Province, Kenya" by EXP Agency, Kenya. Out of the 257 households, 236 were used for analysis after removing surveys that reported filtering more than 70 liters per day or did not have a primary user available to answer questions. By the end of the site visit, the verification team had visited 29 households that had been surveyed by VF/EXP, and 33 unsurveyed

Table 6: Assessment of quality criteria for the third verification report

Source: Own illustration based on Vestergaard Frandsen (2014)

The survey approach and methodology was conducted based on the new GS guidelines for monitoring usage rates of water filters (Gold Standard 2014). The guidelines contain 6 topics that are mandatory to be tested. Generally, it can be concluded that those 6 topics have been tested by the survey methodology and that the survey was in line with the guidelines.

However, parts of the methodology employed by VF are not free of bias. This is the case for the following points:

- Removed households that did not have filtered water in the storage container within the last 24 hours
- Removed households that filtered less frequently than once every two weeks
- Removed households that do not clean the pre-filter or backwash at least once every two weeks

The above points can only be answered through self-reporting. And answers can therefore be heavily influenced by courtesy and recall bias. But it has to be noted that they are in line with the GS guidelines for assessing usage rates for HWFs.

The GS guidelines require that "the survey should not be conducted immediately after capacity building/awareness programs in the target households" (Gold Standard 2014). The VR report says that a "volunteer mobilisation took place in October-November 2013 focused on the correct use of safe storage containers" (ERMCVS 2014b). It is not clear in the VR, if this means that this education campaign was indeed run before the surveys. If it had run before, the survey would not have been in line with the GS guidelines.

In the VR report it is mentioned by the DOE that sub-county coordinators that conduct household visits to carry out checks and repairs of the Lifestraw filters as part of their daily work activities, were also conducting household visits during the survey. The DOE concludes that the independence of the person could be at risk and "the sub-county coordinator conducting the survey in his/her own sub-county in the future" should be avoided. Having said so, it needs to be noted that the GS guidelines do not specifically address the issue of potential adverse incentives of surveyors.

It needs to be concluded that the VF-MR3-Survey has been conducted in line with the GS guidelines and has been verified on this basis. However, from a rigorous and conservative point of view, the quality of the survey and its results could have been improved by limiting the space for recall and courtesy bias and building the survey more on observations that can be objectively made at the point in time of the household visit. The methodology presented in Table x above would be an ideal blueprint for this purpose.

3.4 Stanford study

Usage rate	61.4% (November 2011), 52.1% November 2012, 18.6% May 2013 – November 2013
Survey period	November 2011, November 2012, May 2013 – November 2013
Characteristics of the sample population	Households with pregnant women/caregivers that obtained a LSF under the Vestergaard GS886; 27 rural villages in three counties covered by the programme (Kakamega, Bungoma, and Vihiga)
Sample size	453 (November 2011), 374 (November 2012) and 4041 (May 2013 – November 2013)
Approach	Randomized control trial (no details available) Self-reporting and observations
Methodology	Field staff asked respondents to fetch a cup of water the way they normally would for a young child, then observed from where the respondent obtained the water and how it was stored and extracted. Field staff inquired if anyone in the household "had done anything to make the water

	less cloudy or safer to drink", and if so, what method was used (without prompting on specific water treatment methods). If the respondent did not report a water treatment method, the field staff asked if the respondent ever treats drinking water and to list all methods used. Field staff questioned if the household had received a LSF. If the household reported receiving a filter, we observed if the filter was present, hanging on the wall, looked unused (e.g., visible dust), and contained water or moisture. Field staff asked if the filter was working and if there were any issues that prevented use. Finally, respondents reported if and when a representative from the Carbon for Water program had most recently visited their home to promote the LSF.
Surveyor	Non-profit organization Innovations for Poverty Action (IPA) in Kenya
Quality control	None mentioned

Table 7: Assessment of quality criteria of the Stanford study

Source: Own illustration based on Pickering et al. (2016)

The sample population has similarities and differences with the total population of households that received LSF. Firstly, the sample contains exclusively pregnant women or caregivers. In this sense the sample group is a sub-group of 80% of the households (women) that received a LSF. Therefore, one can determine the sample group as generally representative. The following point could even hint towards an increased use of LWF in this sample (which would make the results conservative). It can be assumed that pregnant women would exercise special care with their diet and possible sources of contamination as they are aware they are carrying a baby. Secondly, the Stanford surveys have only been conducted in 2 (1st survey), 2 (2nd survey) and 3 (3rd survey) out of 5 project areas of Vestergaard GS886 and 2, 2 and 8 out of 32 sub-counties of Vestergaard GS886. However, especially for the 3rd Stanford survey, it should be assumed that the Stanford results are similarly representative to the results of VF due to the sufficient sample size in the Stanford study and due to the fact that a quarter of the sub-counties have been sampled.

The survey approach and methodology mainly relies on a very balanced set of observations and self-reporting. When comparing the methodology to the GS guidelines it has to be noted that 2 out of 6 topics are not addressed by the methodology. Firstly, it is not checked, if the household stores the filtered water (if not they would be considered a non-user). Secondly, it is not asked how often the household uses the water filter (low frequency users would have been considered non-users). If those questions would have been addressed in the survey, it can be assumed that the usage rate would have been even lower that the already considerably lower rates compared to the VF usage rates.

One of the key strengths of this study seems to be that the survey has been conducted by a credible NGO specialised in data collection in collaboration with research institutes (IPA 2017) that seems independent and had no direct incentive in over- or underreporting of the usage rate. However, Gold Standard says that IPA had a conflict of interest due to the following reasons (Gold Standard 2017c):

- "IPA had raised several million dollars from prominent donors for its own Dispensers for Safe Water Program program (DSW)" using chlorine dispenser technology
- "At an organizational level IPA had a strategic interest in the success of DSW due to the amount of money raised, the prominence of the donors, and the fact that it was a key program for IPA"
- In 2012, IPA was contracted by Impact Carbon as the implementing and distribution partner for its centrally located chlorine dispenser project GS966 which took place in the same boundary as Vestergaard GS886. And in 2013, IPA created a separate legal entity, Evidence Action, to spun-off the part of the organisation involved in the implementation of GS966. Gold Standard therefore claims that IPA "had a strategic interest in that project's [comment by the author: GS 966's] success through Evidence Action".

But in contradiction to the above statement that "IPA has a conflict of interest", Gold Standard in the same report acknowledges that "there was likely no actual conflict of interest at the organizational or project management level based on the following factors: (1) IPA's large, de-centralized structure, (2) rigorous quality controls, (3) efforts to keep the local enumerators objective (i.e., the researchers trained and managed the data collectors – not IPA); and (4) the fact that the researchers' activities, office space, and employees were wholly separate from IPA's other activities, including any other chlorine dispenser program." Gold Standard rather sees an indirect conflict of interest by concluding that "it is impossible to ignore the bigger strategic and financial picture for IPA, which may have created bias at the enumerator or household level. Specifically, households may have had previous exposure to IPA and chlorine dispensers; enumerators introduced themselves as IPA employees, which may have created survey bias; and, as local community members, enumerators may have had their own biases against Lifestraw water filters." Although it is not possible to completely rule out this potential bias of the enumerators, this Gold Standard argument seems relatively weak. Especially, when taking into account that, as shown above, volunteers recruited by Vestergaard Frandsen which received a nominal fee as payment and sub-county coordinators employed by Vestergaard Frandsen carried out the surveys in Vestergaard GS886 which creates a very clear and direct conflict of interest.

One weak point of the Stanford study is that there is no independent quality control over the results. However, it must be questioned if this is necessary as the study has been guided by independent scientists and as the data has been collected by an independent NGO. Additionally, it needs to be

taken into account that the main quality control (survey among 257 households) in Vestergaard GS886 was undertaken by EXP Agency, a firm that was hired by Vestergaard Frandsen for conduction awareness-raising campaigns and stakeholder consultations during the project implementation.

3.5 Discussion of findings

Comparison across all monitoring periods

Vestergaard GS886		Stanford
Monitoring Period	**Usage rate**	**Usage rate**
01.06.2011 - 30.11.11	91,13%	61,40%
01.12.11 - 31.10.12	74.98%	52,10%
01.11.2012 - 31.01.2014	80,46%	18,60%

Table 8: Comparison of the usage rate across all reviewed verification reports

Source: Own illustration

For each monitoring period a large discrepancy in the usage rates can be observed. However, it has to be taken into account that the sample sizes in the Stanford study were considerably lower. Using an accepted formula in statistics one can calculate the relative precision of the results obtained given by a certain sample size (the higher the sample size the higher the relative precision and the lower the range of the usage rate) (Feige et al. 2012). We have calculated the rage of usage rates for VF and Stanford for all three monitoring periods. Afterwards, we have determined the lowest/conservative difference between the usage rates obtained in the two studies in all periods.

Vestergaard GS 886: Range of usage (normalised based on sample size)				
Monitoring Period	**Sample size**	**Relative precision of results**	**Range of usage**	
01.06.2011 - 30.11.11	19430	0,7%	90,43%	91,83%
01.12.11 - 31.10.12	20220	0,7%	74,28%	75,68%
01.11.2012 - 31.01.2014	16313	0,7%	79,76%	81,16%
Stanford: Range of usage Range of usage (normalised based on sample size)				
Monitoring Period	**Sample size**	**Relative precision of results**	**Range of usage**	
01.06.2011 - 30.11.11	453	5,0%	56,40%	66,40%
01.12.11 - 31.10.12	374	5,0%	47,10%	57,10%
01.11.2012 - 31.01.2014	4041	1,5%	17,10%	20,10%

Table 9: Range of usage for the reviewed reports

Source: Own calculations

The results still show a significant difference in the usage rates. Especially in the 3rd monitoring period. We therefore discuss the findings for each MP separately.

1st monitoring period

VF survey had significant flaws that have been addressed above:

- Leading questions
- Only self-reporting
- Done shortly after awareness-raising campaigns
- Not random selection?
- 91.13% likely not a representative usage rate

From a scientific point of view, the approach and methodology by Stanford seems more appropriate compared to the survey undertaken by VF. In absence of any other data **61.40%** as determined by the Stanford study seems a more appropriate usage rate for this period.

2nd monitoring period

Original survey had similar flaws as the one conducted during 1st monitoring period. Usage rate of 74.98% recommended by Stanford is based on a quantitative post-processing of the survey approach and method used by VF and the data obtained through it. The post-processing had certain limitations (e.g. respondents in the survey could not be asked "better" questions, etc). Some of the expert comments suggest that, if other approach and methods would have been used, the usage rate could also have been lower than 74.98%.

In the absence of any other data **52.10%** usage rate from Stanford seems to be more appropriate usage rate.

From a good governance point of view it was not appropriate by the Gold Standard to mandate the 74.98% based on a post-processing of data surveyed by VF, as methodology and PDD require sampling to determine the usage rate. In our point of view, Gold Standard should have asked VF to conduct a new survey for the 2nd monitoring using the new guidelines.

3rd monitoring period

The large discrepancy in usage rates between Vestergaard GS886 and Stanford cannot be explained with large differences in how the surveys have been carried out. However, it needs to be taken into account that a large number of the surveyors could have had an interest to report high usage rates.

However, this is not prevented by the GS guidelines. Additionally, the survey was probably conducted shortly after an education campaign (which would not be in line with the GS guidelines for HWF projects) and should be investigated.

What also can be observed is that the Stanford study has been conducted scientifically sound from the first survey onwards. The results show a decreasing usage rate. The study shows that although LSFs exist in almost all households that once received a filter (still 93.6% after 3 years), it is not used anymore by an overwhelming majority of the households. These findings are backed-up by a comprehensive investigation based on observations into why the filters are not being used anymore (which is the preferred method for conduction a survey over self-reporting).

Households that produced filter			
Total number	421	354	3616
Filter not working	20 (4.8)	77 (21.8)	1839 (51.0)
Don't know	0	0	8
Filter hanging on wallΦ	393 (93.4)	340 (96.9)	3360 (93.8)
Observation not possible	0	3	33
Moisture in filter reservoirΦ	136 (32.3)	89 (25.2)	438 (12.2)
Observation not possible	0	1	18
Filter has signs of non-use (*e.g.* dust)Φ	299 (71.0)	228 (64.8)	2963 (82.3)
Observation not possible	0	2	17

Φ Direct observation by field staff

Table 10: Households that produced filter

Source: Pickering et al. (2016)

It must be therefore assumed that the difference in the usage rates results from the actual setting in which the interviews were conducted, the room for bias (e.g. courtesy and temporal), the amount of training that the interviewers received in order to perform the interviews properly, perception of respondents on what they would be expected to say, etc. It is therefore impossible to objectively determine the source of the large discrepancy between the usage rates obtained.

In the absence of any other data, it should be assumed that the usage rate is **18.60%**.

It is also noteworthy that the Stanford study determines which percentage of the filters showed moisture. This is a very strict criteria, but also the most plausible indicator for defining recent usage. When applying only this criteria, the usage rate would only be 12.2%. However, in Vestergaard GS886 every household that uses the LSF every two weeks is defined as a user (ERMCVS 2012). The Berkeley study finds that “moisture dries quicker than common usage patterns for the filter. The filter may be wiped or dried in a few hours when the household filters twice per week, so enumerators would not be able to use those signs.” and that therefore a moist filter is not an appropriate indicator for determining use based on the definition of a user in Vestergaard GS886 (Graham et al. 2013). If a

single usage within two weeks is a sufficient indicator for the household to count as a user and allow the project to generate carbon credits for this household, is of course debatable. When assuming that a household of five members needs 17.5 – 35.0 L[1] of water per day for drinking and cooking, it does not seem reasonable to assume that this amount would be produced only once in 2 weeks. The corresponding volume of water storage that would need to be filled with LSF and stored at one time would be 245 or 490 L. It therefore seems much more likely that the households fill the water storage every day or every second day using LSF.

4 Review of the approach and methodology and the results for calculating the programmes' wrongly issued VERs by atmosfair

We have calculated scenarios for the overissued VERs. In doing so, we have applied different usage rates. The results are displayed in the following table and are broadly in line with previous atmosfair calculations.

Scenario description	Vestergaard Frandsen		**Scenario 1: Stanford usage**			**Scenario 2: Stanford usage - moist filter only indicator**		
Monitoring Period	Usage rate	Issued VERs	Usage rate	Corres-ponding VERs	**Overissued VERs**	Usage rate	Corres-ponding VERs	**Overissued VERs**
MP 1	91.13%	1,351,102	61.40%	910,362	**440,740**	32.30%	478,904	**872,198**
MP 2	74.98%	1,701,563	52.10%	1,182,334	**519,229**	25.20%	571,878	**1,129,685**
MP 3	80.46%	1,419,458	18.60%	328,119	**1,091,339**	12.20%	215,218	**1,204,240**
Total		4,472,123		2,420,816	**2,051,307**		1,265,999	**3,206,124**

Table 11: Overissued VERs in different usage rate scenarios compared to Vestergaard Frandsen rates

Source: Own calculations

It can be seen that the overissued VERs are around 2,051,000 (+/- 5%[2]). If one takes into account that the user definition in Vestergaard GS886 is very lenient (filtering up to every 2 weeks) which

[1] 7.5 Liter per person per day is the maximum amount of water need eligible for generating VERs in Vestergaard GS886

[2] Taking into account the relative precision of 5% in Table 9

prevents using moisture in the filter as an observation as a strict indicator, and instead using moisture as the only indicator, the overissued VERs reach up to a volume of 3,200,00.

5 Conclusions

Based on the above analysis, it can be concluded that very likely an overissuance for the Vestergaard GS886 has occurred. This conclusion is based on the observation that relative to the Stanford study, the survey undertaken by Vestergaard Frandsen had the following considerable flaws that very likely led to an overreporting of usage rates:

- a direct conflict of interest of the surveyors
- a large potential for courtesy bias during the interviews
- at least two of the three surveys have been carried out directly following awareness-raising campaigns (possibly this was also the case in the third survey)
- two of the three surveys have been conducted with leading questions

It therefore has to be concluded that the Stanford study is more representative of the real usage rates in Vestergaard GS886 and that therefore the resulting amount of overissued VERs range between 2,051,000; and potentially even 3,200,00.

The Gold Standard sees itself as the prime and "gold" standard for carbon credits. This is for example documented by the following statements on the Gold Standard website (Gold Standard 2017):

- "the Gold Standard has always had rigorous safeguards in place to ensure projects deliver real (additional) GHG emissions reductions"
- "Since it was established in 2003 by WWF and other international NGOs, Gold Standard has set the benchmark for best practice climate initiatives, ensuring they contribute to sustainable development as well as climate mitigation."

This claim is also echoed by the carbon market in which the reputation of the Gold Standard is seen as excellent. It is apparent that the grievance process and its side-effects could result in an erosion of this reputational status. In this regard it is particularly striking that the Gold Standard did not make an attempt right at the start of the complaints about the alleged overreported usage rates to proactively end this debate. For example, one solution could have been not to issue the VERs for the 2nd monitoring period until usage rate have been established in a scientifically sound manner. The Gold Standard could have asked for a truly independent survey on the usage rates applying

best–practice independent sampling procedures. The approach finally taken by Gold Standard (to allow the verification and issuance to happen based on a usage rate that contrary to the Gold Standard rules had not been determined by sampling) is problematic when considering its core principles.

References

Atmosfair gGmbH (2016b): Annex to the Launch of Grievance Procedure with regard to the Project Sustainable Deployment of the LifeStraw Family in Rural Kenya (GS886).

Atmosfair gGmbH; Germanwatch; European Business Council for Sustainable Energy (2013): Comments of Atmosfair and Germanwatch on the LifeStraw Project, 2nd verification.

CDM Executive Board (2011): Project Design Development Form: Sustainable Depolyment of the LifeStraw Family in Rural Kenya.

Det Norske Veritas (2010): Gold Standard Validation Report: Sustainable Deployment of the LifeStraw Family in Rural Kenya.

ERMCVS (2012): Gold Standard Verification Report: Sustainable Deployment of the LifeStraw Family in Rural Kenya.

ERMCVS (2013): Gold Standard Verification Report: Sustainable Deployment of the LifeStraw Family in Rural Kenya.

ERMCVS (2014a): Gold Standard Verification Report: Sustainable Deployment of the LifeStraw Family in Rural Kenya. 2nd Verification.

ERMCVS (2014b): Gold Standard Verification Report: Sustainable Deployment of the LifeStraw Family in Rural Kenya. 3rd Verification.

Feige, S.; Marr, M. (2012): Sampling manual: A guide to sampling under the CDM with special focus to PoAs. First Edition. KfW Bankengruppe.

Gold Standard (2010): Indicative Programme, Baseline and Monitoring Methodology for Improved Cook-Stoves and Kitchen Regimes.

Gold Standard (not date): Engagement Process: Tom Heinemann Climate Documentary.

Gold Standard (no date): Gold Standard Secretariat responses to comments provided by Atmosfair and Germanwatch on the GS886 – LifeStraw deployment in Kenya – 2nd verification.

Gold Standard (no date): Gold Standard Local Stakeholder Consolidation Report.

Gold Standard (2014): Guidelines for Carrying Out Usage Surveys for Projects Implementing Household Water Filtration Technologies.

Gold Standard (2016): Report on Grievances related to Vestergaard LifeStraw Project.

Gold Standard (2017a): Challenges to Usage Rates of LifeStraw Water Filters. Geneva, Switzerland.

Gold Standard (2017b): Risk Assessment Investigation Plan for GS886: Sustainable Deployment of the LifeStraw Family in Rural Kenya.

Gold Standard (2017c): GS886 Lifestraw Water Filter Project – Non Conformity overview and recommendations.

Gold Standard (2017d): Our work. Our principles. Gold Standard website. URL: www.goldstandard.org/our-work/our-principles-process; 27.09.2017

Graham, J.; Kaur, M.; Pennise, D. (2013): Assessment of Usage Methods of GS886 Sustainable Deployment of the LifeStraw Family in Rural Kenya. Berkeley Air Monitoring Group.

Heinemann, T (2013): The Carbon Crooks. Film documentary. URL: http://carboncrooks.tv/; 17.09.2017

Innovations for Poverty Action [IPA] (2017): About. What we do. IPA website. URL: www.poverty-action.org/about/what-we-do: 17.09.2017

Letter from Dr. Dietrich Brockhagen (Atmosfair gGmbH) to Mr. Goyal (Gold Standard); 26.03.2013.

Letter from Dr. Dietrich Brockhagen (Atmosfair gGmbH) to Marion Verles (Gold Standard). Launch Grievance Procedure on false CO_2-Emission Compensation: Sustainable Deployment of the LifeStraw Family in Rural Kenya (GS886); 14.04.2016.

Letter from Marion Verles (Gold Standard) to Dr. Dietrich Brockhagen (Atmosfair). Launch Grievance Procedure on false CO_2-Emission Compensation: Sustainable Deployment of the LifeStraw Family in Rural Kenya (GS886); 12.06.2016.

Letter from Dr. Dietrich Brockhagen to Marion Verles. Grievance regarding Sustainable Deployment of the LifeStraw Family in Rural Kenya (GS886) Project. Reply to Answer of the Gold Standard; 15.08.2016.

Pickering, A.J.; Arnold, B.F.; Dentz, H.N.; Colford Jr, J.M.; Null, C. (2016): Climate and Health Co-Benefits in Low-Income Countries: A Case Study of Carbon Financed Water Filters in Kenya and a Call for Independent Monitoring. In: Environmental Health Perspectives, 125; p. 278-283.

Starr, K. (2011): Thirty Million Dollars, a Little Bit of Carbon and a Lot of Hot Air. Stanford Socila Innovation Review. Available at https://ssir.org/articles/entry/thirty_million_dollars_a_little_bit_of_carbon_and_a_lot_of_hot_air Accessed September 2017.

Vestergaard Frandsen (2012): GS0086 Sustainable Deployment of the LifeStraw Family in Rural Kenya, Verification Report, Verification 1.

Vestergaard Frandsen (2013): GS886 Sustainable Deployment of the LifeStraw Family in Rural Kenya, Verification Report, Verification 2 (final version).

Vestergaard (2014): GS886 Sustainable Deployment of the LifeStraw Family in Rural Kenya, Verification Report, Verification 3.