Demand Response Evaluation, Cost-Effectiveness and System Planning

Nik Schruder, Ontario Power Authority, Toronto, ON
Josh Bode, MPP, Freeman, Sullivan & Co, San Francisco, CA

ABSTRACT

This paper summarizes the historical performance and reliability of aggregator demand response resources in Ontario. Since 2008, Ontario Power Authority (OPA) has operated a contractual demand response (DR) program which offers reserve payments to participants in return for providing load reductions when called upon. The program primarily targets aggregators but also allows direct participation by customers that can reduce demand by more than 1 MW. It currently has 400 MW of contractual resources delivered by three aggregators and four direct participants at over 500 facilities. This paper focuses on aggregator performance. We present the performance and reliability results from 44 contractual DR activations over a five-year period from 2008 to 2012. We use the historical data from the 44 events to explore trends and variability in response patterns based on industry, customer size and other factors. The results from this paper can help guide decisions regarding performance of DR resources that provide a guaranteed reduction in load.

Introduction

Aggregator programs with a focus on commercial and industrial (C&I) customers are wide spread across North America. According to FERC's 2012 *Assessment of Demand Response and Advanced Metering*, reported C&I demand reduction capability exceed 27,000 MW, much of which is delivered through aggregators which either directly participate in wholesale electricity markets or contract with utilities (FERC 2012). Not all aggregators publicly report their demand reduction capability, but those who do report having over 12,000 MW of DR resources.

Despite the large amount of aggregator DR resources, there is limited publicly available data on their reliability and performance, particularly on performance under actual dispatch conditions (versus self-scheduled tests). Data on the performance of aggregator programs is publicly available for both California and PJM Interconnection. PJM reports performance for its load management resources on an annual basis. Between 2008 and 2012, PJM called nine load management events (PJM 2012). However, PJM load management events are localized—typically specific zones are dispatched—and include resources besides those operated by aggregators. California also requires each investor owned utility to evaluate and publicly report performance of DR programs (Braithwait 2009, Braithwait 2010, Braithwait 2011, Bode 2012, Braithwait 2012). Roughly 82 percent of these aggregator DR resources are direct contracts between aggregators and utilities that get dispatched one or two times per year. The performance of these programs is reported annually and, to date, there have been no published multi-year comparisons of performance that we are aware of.

Analyzing performance of OPA's aggregator programs for 44 events over five years allows us to explore the reliability of aggregator programs and variability in response patterns based on industry type, customer size and other factors. We also compare performance based on both settlement baselines and evaluation results. The current baseline method for the OPA's aggregator program has been shown to be biased and, in aggregate, overstates demand reductions by approximately 20 to 25 percent. Despite the settlement baseline bias, performance relative to settlement baselines is important since it is the metric by which aggregators are compensated.

There are three main aspects to aggregator performance: the ability to build DR resources, the ability to ensure the resources remain available, and the ability to deliver expected DR resources. The ability to build resources is analogous to commitments to build generators. Not all generators scheduled to be built are built on time or even built at all. Likewise, DR aggregators can meet, exceed, or fall short of commitments to build new DR resources according to a pre-specified schedule. The ability to build resources according to schedule is not typically factored into performance reliability of either generators or DR resources. However, it has real implications for long-term system planning. Once DR resources are built, the key question is how reliably they perform relative to the reductions expected by operators. Reliability includes two components: pre-announced non-performance, typically due to facility maintenance of large electricity customers; and deviations between DR resources scheduled and delivered during actual activations. These differences can be attributed to shortfalls or over delivery by aggregators, or structural flaws such as bias in the settlement baselines.

The remainder of this paper is divided into sections that provide background, describe the methodology, present results, and summarize conclusions. Unless otherwise indicated, the tables and graphs present results for OPA's aggregator resources and exclude large customers that participate without an aggregator.

Background

OPA's aggregator program, DR-3, has been in operation since 2008, and allows participants and aggregators to enter into contractual agreements for load reductions with OPA. Aggregators and direct participants must commit to a specific load reduction amount for either 100 or 200 hours per year. DR resources must be available from 12-9 pm, June through September, and from 4-9 pm, all other months. Participants also have the option of a shorter summer availability period, from 12-6 pm, for lower incentives, but few customers have elected that option. In exchange for DR load reduction capability, the DR-3 program makes both availability (capacity) and energy payments to aggregators and direct participants. While payments and performance are assessed in aggregate (by settlement account), aggregators must specify the expected demand reduction capability for each individual site. In addition, aggregators must specify for each individual contributor whether DR resources will be delivered through load reduction or behind-the-meter generation. Behind-the-meter generation is allowed in Ontario's DR-3 program, provided the generator meets specific environmental standards.

OPA has discretion regarding the timing and duration of events. Dispatch of DR-3 is coordinated with Ontario's Independent Electricity System Operator (IESO) and typically tied to estimates of the supply cushion—the degree by which available supply resources exceed demand. The resource can be activated for both emergency and economic reasons.

Participants must notify OPA and the IESO of any short-term fluctuations in load reduction capability due to facility maintenance or down time. These days are classified as non-performance days and are analogous to scheduled generator outages. They enable the IESO to better operate the system and schedule alternate resources for those days. Scheduled non-performance leads to reductions in the participant payments. Unscheduled non-performance—failure to meet contractual obligations during events—leads to even larger payment reductions.

Figure 1 summarizes the change in DR resources and number of sites since the program's inception in 2008. For clarity, the figure separately presents aggregator and direct participant resources. By the end of 2012, there were over 500 sites in the program with aggregate DR resource contract to deliver up to 402 MW during summer months. Since program launch, the participant mix and load reduction capabilities have evolved substantially. Initially, the load reduction capability was highly concentrated among direct participants. Over time, aggregator DR resources have grown and now

account for over 80 percent of DR-3 resources. In addition, the mix of participating aggregators has consolidated over time, from eight to three aggregators.

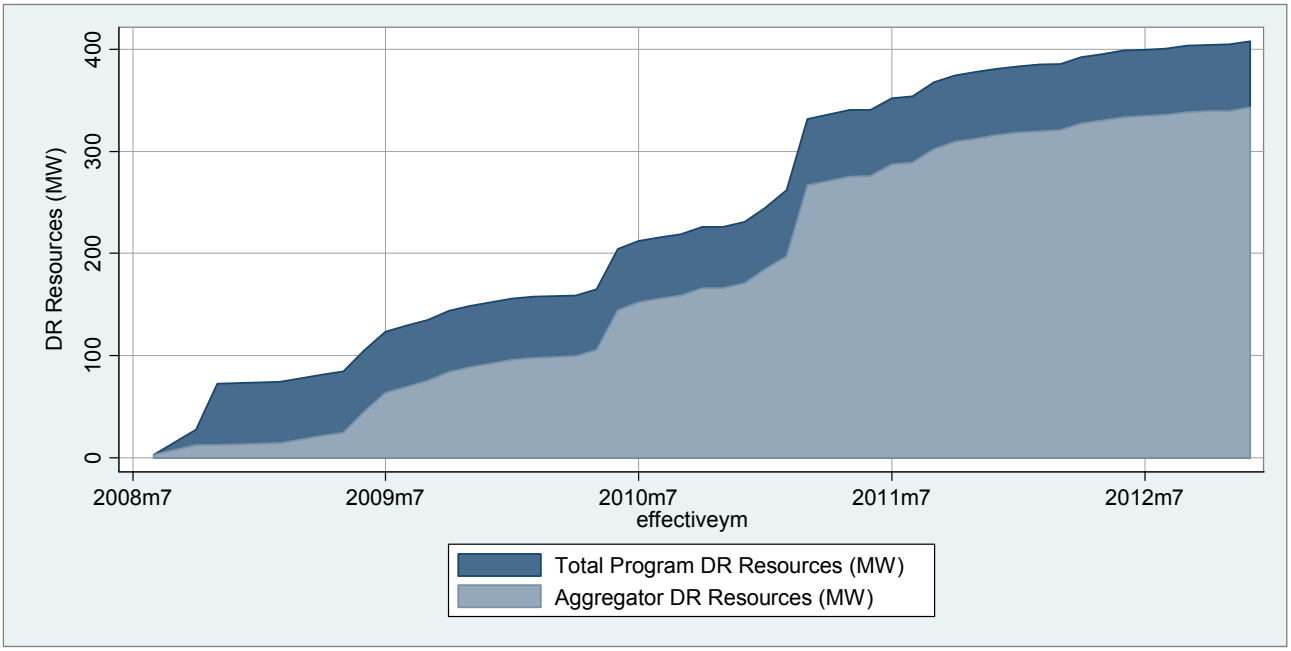

Figure 1: Growth in Program and Aggregator DR Resources 2008-2012

Despite the large number of facilities—528 enrolled through aggregators—most of the aggregator load remains highly concentrated among large electricity customers. Figure 2 shows the concentration of DR resources and program load for OPA's aggregator resources. Customers were ranked based on their stated demand reduction capability and grouped into 10 deciles of 53 customers each. Overall, 10 percent of aggregator contributors (53 out of 528) account for over 70 percent of aggregator DR resources and over 50 percent of the loads. This high concentration of resources has several implications for how well the program performs when dispatched. DR resources, which are highly concentrated among a small set of customers, typically exhibit more variability in performance because a few sites can greatly affect the amount of DR delivered. More diverse resources tend to exhibit less variability.

Figure 3 visually depicts the program's concentration of customers, peak load, and demand reduction commitments across industries. The industry groups are ordered based on the aggregate demand reduction commitment. For example, contributors in the "Industrial Tool/Metal Works" category account for 6.6 percent of customers but account for 14.2 percent of the program load and 22.4 percent of the program demand reduction commitments. The "Timber, Pulp & Paper" sector accounts for less than 2 percent of customers, but accounted for 12.1 percent of the program's peak load and 22.4 percent of the demand reduction commitments. In other words, these customers not only use more electricity than the average contributor, but are expected to reduce a greater share of their load than the average customer. The reverse is true for "Offices, Hotels, Banks, and Professional Services." They account for 36.9 percent of the program's customers, but represent only 17.4 percent of the load and 4.1 percent of the demand reduction commitments. In other words, these customers use less electricity than the average contributor, and also reduce a smaller share of their loads than the average customer.

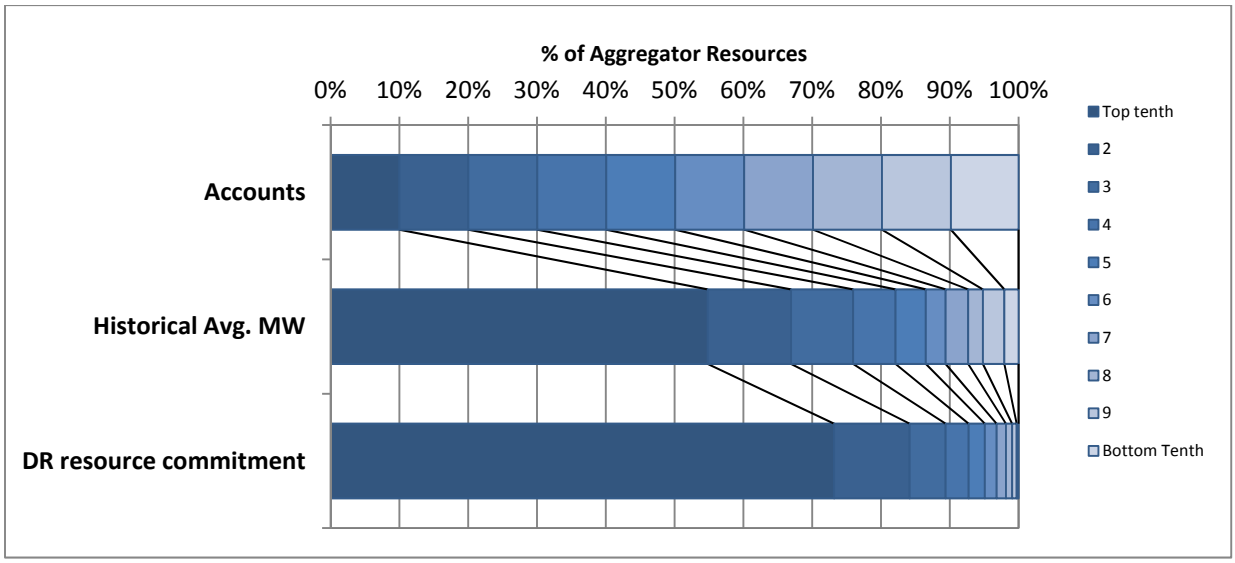

Figure 2: Concentration of OPA's Aggregator DR Resources (December 2012)

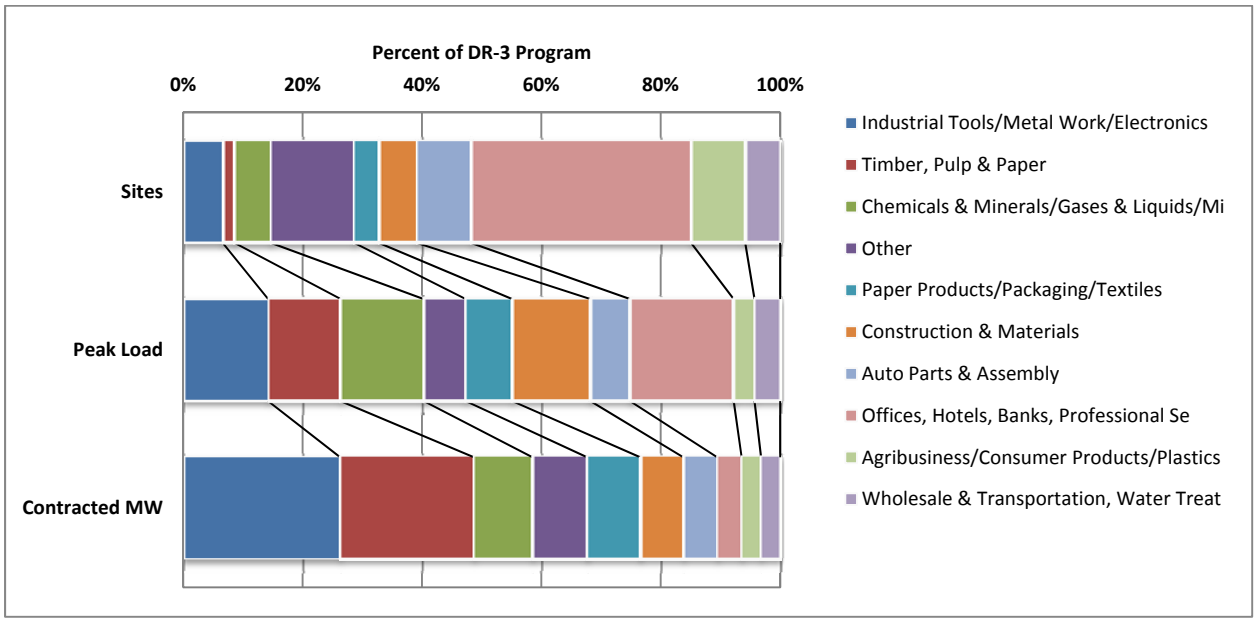

Figure 3: Concentration of Customers, Loads and Demand Reduction Commitments by Industry (December 2012)

A few additional noteworthy facts about the composition of aggregator resources are:

- 45 percent of aggregator DR resources, or 154 MW, participated in the 200-hour dispatch option. Coincidentally, 45 percent of individual contributors are assigned to the 200-hour dispatch option. The remainder of individual contributors and DR resources are assigned to 100-hour dispatch options.
- 85 of individual contributors supply 27.3 MW of aggregator DR resources using behind-the-meter generation.

From 2008 to 2012, OPA dispatched the program 44 times. In each instance, customers signed up for the 200-hour option were dispatched, while customers on the 100-hour option were activated more sparingly and were dispatched 31 times. In total, both options were dispatched jointly 31 times.

Figure 4 summarizes the weather and Ontario system load conditions when DR-3 was dispatched. Table 1 shows the distribution of events by year, month, day of week, and start time for each option.

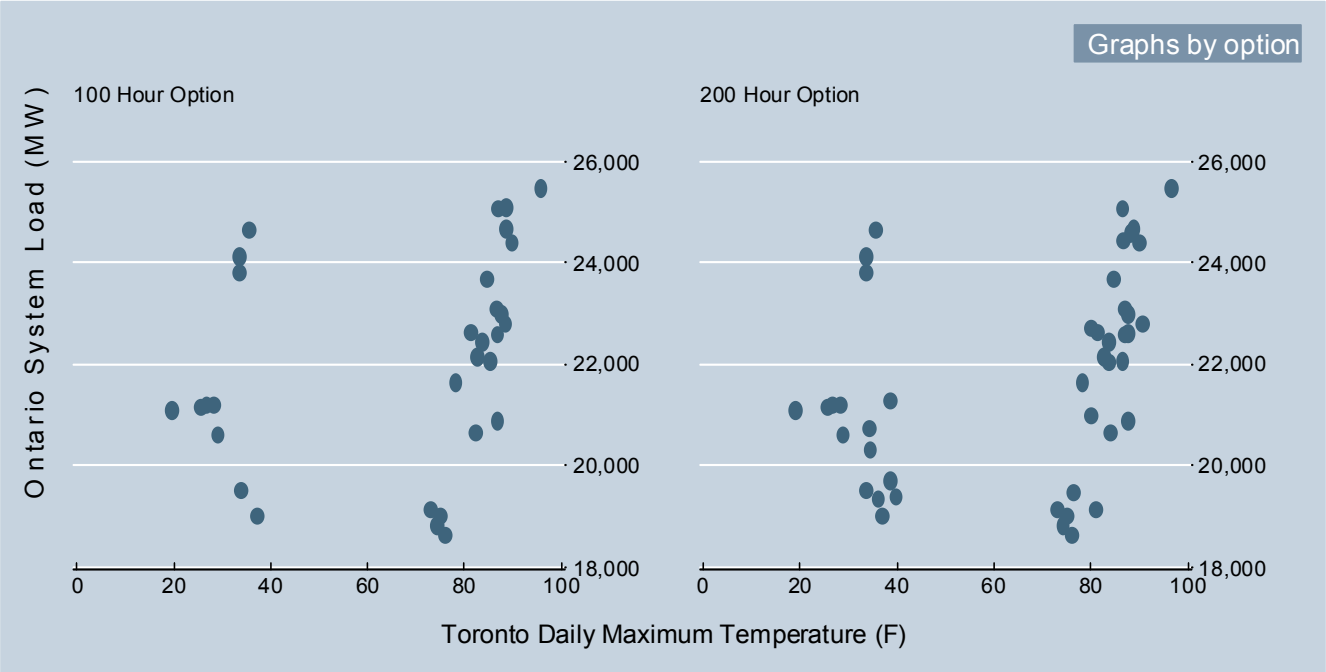

Figure 4: Event System Load and Weather Conditions (2008-2012)

Table 1: Distribution of Event Conditions by Year, Month, Day of Week, and Start Time

Year	100 hour option	200 hour option
2008	8	14
2009	4	6
2010	6	8
2011	8	11
2012	5	5
Total	**31**	**44**

Month	100 hour option	200 hour option
May	2	2
June	4	6
July	6	8
August	5	6
September	9	11
October	0	2
November	4	8
December	1	1
Total	**31**	**44**

Day of Week	100 hour option	200 hour option
Monday	4	10
Tuesday	8	9
Wednesday	8	13
Thursday	8	9
Friday	3	3
Total	**31**	**44**

Event Start Time	100 hour option	200 hour option
12:00 PM	2	2
1:00 PM	5	7
2:00 PM	9	10
3:00 PM	8	11
4:00 PM	6	9
5:00 PM	1	5
Total	**31**	**44**

For context, Ontario system loads peaked on July 21, 2011 at 25,450 MW during the 2008 to 2012 period. On that day, Toronto temperatures peaked at 97°F (36°C). Typically, Ontario loads peak during hot, humid summer days. However, the province also experiences high system loads during winter months due to the use of space heating. The system loads in some of the shoulder months, such as November, can lead to shortages in available supply capacity because electricity is also used for heating in many parts of Ontario. As a result, aggregator resources have been dispatched not only during

summer months but throughout the year. Aggregator resources have been dispatched Monday through Friday and at varying start times. The wide number of events and wide variation in event conditions prove useful in assessing the reliability and performance of aggregator resources.

The remainder of this paper presents results for sites associated with aggregator resources only. It excludes direct participants, sites that enrolled after the 2012 summer, and a small number of sites (<5%) for which we were unable to incorporate 2012 results.

Methodology

We estimated load reductions for each individual contributor, using within-subject methods, and subsequently aggregated results across sites. For each year, the final impacts were estimated through regression methods using data from event and non-event days and available pre-enrolment data. We also assessed the accuracy of different day matching baseline methods in each evaluation year. At a fundamental level this approach uses information from non-event days (which can be thought of as control days) to estimate the amount of electricity customers would have consumed had aggregator resources not been dispatched (the reference load). The program impact is the difference between the reference load and actual consumption during DR event activations.

We first tested how accurately different models estimated loads during event like days (proxy events) when aggregator resources were not dispatched. Proxy event-day impact estimates should be insignificant and centered around zero because, in fact, aggregator resources were not dispatched.[1] The estimates were produced out-of-sample, meaning the proxy events and actual event days were withheld from the model development and did not inform the development of the counter factual. To assess accuracy, we compared actual and estimated loads for those days. The model with least error was then applied to estimate impacts for actual event days. OPA's annual evaluations of demand response programs provide more details regarding the methodology used to develop impact estimates.

While the evaluation attempts to produce unbiased results that are measured as precisely as possible, it is important to keep in mind that impact estimates for each event day have uncertainty. The confidence bands vary for each event based on the number of sites and the customers mix. The confidence bands for the average event are narrower since positive and negative errors during individual event cancel each other out.

Another key limitation within subject methods is that event days typically are systematically different than non-event days; event days tend to have more extreme weather and/or higher system loads. In general, it is preferable to rely on methods that use both non-event-day information and a comparable control group (e.g., difference-in-differences). Methods that rely on control groups do not require extrapolating from non-event to event-day conditions. However, a control group could not be developed because OPA cannot include non-participant data as part of the evaluation due to confidentiality and privacy concerns. In addition, control groups cannot easily be applied in this context due to the concentration of resources among larger industrial sites and the fact that different resources are dispatched for each event.[2]

The primary focus of the analysis was to estimate actual load reductions relative to the contracted and scheduled load reductions (the realization rate). The actual causes for differences between contractual and delivered aggregator resources can only be assessed through descriptive analysis. Potential reasons for differences between contractual and delivered reductions include

[1] The programs were not activated during proxy events and customers do not otherwise have an incentive to alter their normal behavior.

[2] This occurs within each option due to continued program growth and because aggregators and direct participants can schedule non-performance days in advance.

systematic biases in the settlement baselines, variation in scheduled non-performance, measurement error, and aggregator non-performance.

Results

Analyzing performance of aggregator programs for 44 events over four years allows us to address several research questions:

- What is the overall performance and volatility of the aggregator resources?
- Does performance improve with program growth and experience?
- Does volatility in event-day performance decrease with program growth and experience?
- What share of the performance gap is due to error in settlement baselines? How do aggregators perform based on the settlement baselines that directly affect them?
- Does performance vary substantially by industry, customer size or amount of dispatch experience?

The pattern of reliability for aggregator resources differs from that of generators. For generation resources, scheduled or forced outages often mean the resource is entirely unavailable. In contrast, the effect of pre-announced non-performance is more nuanced for DR aggregator resources. While aggregator scheduled and delivered resources vary and sometimes fall short of nameplate capacity, the shortfalls are partial.

Figure 5 summarizes the overall performance of aggregator resources in Ontario from 2008 to 2012. It compares side-by-side the delivered, scheduled, and contracted DR resources. The results are presented separately for the 100- and 200-hour options and for instances when both options were jointly dispatched. The delivered reductions in the graph are based on evaluation results and not on the settlement baselines. As we discuss later, it highlights the importance of assessing accuracy of settlement rules in advance of signing multi-year contractual agreements. The gap between the delivered and scheduled DR resources reflects deviations between DR resources expected by the system operator and resources delivered during actual activations. A substantial portion of this difference is explained by systematic bias in the settlement baselines. The difference between delivered and contracted resources is similar to comparing generator reliably against nameplate capacity. The ratio between delivered and contracted resources reflects the realization rate. Across all five years on average 91 percent of contracted resources were scheduled and 91 percent of scheduled resources were delivered, producing a realization rate of 83 percent.

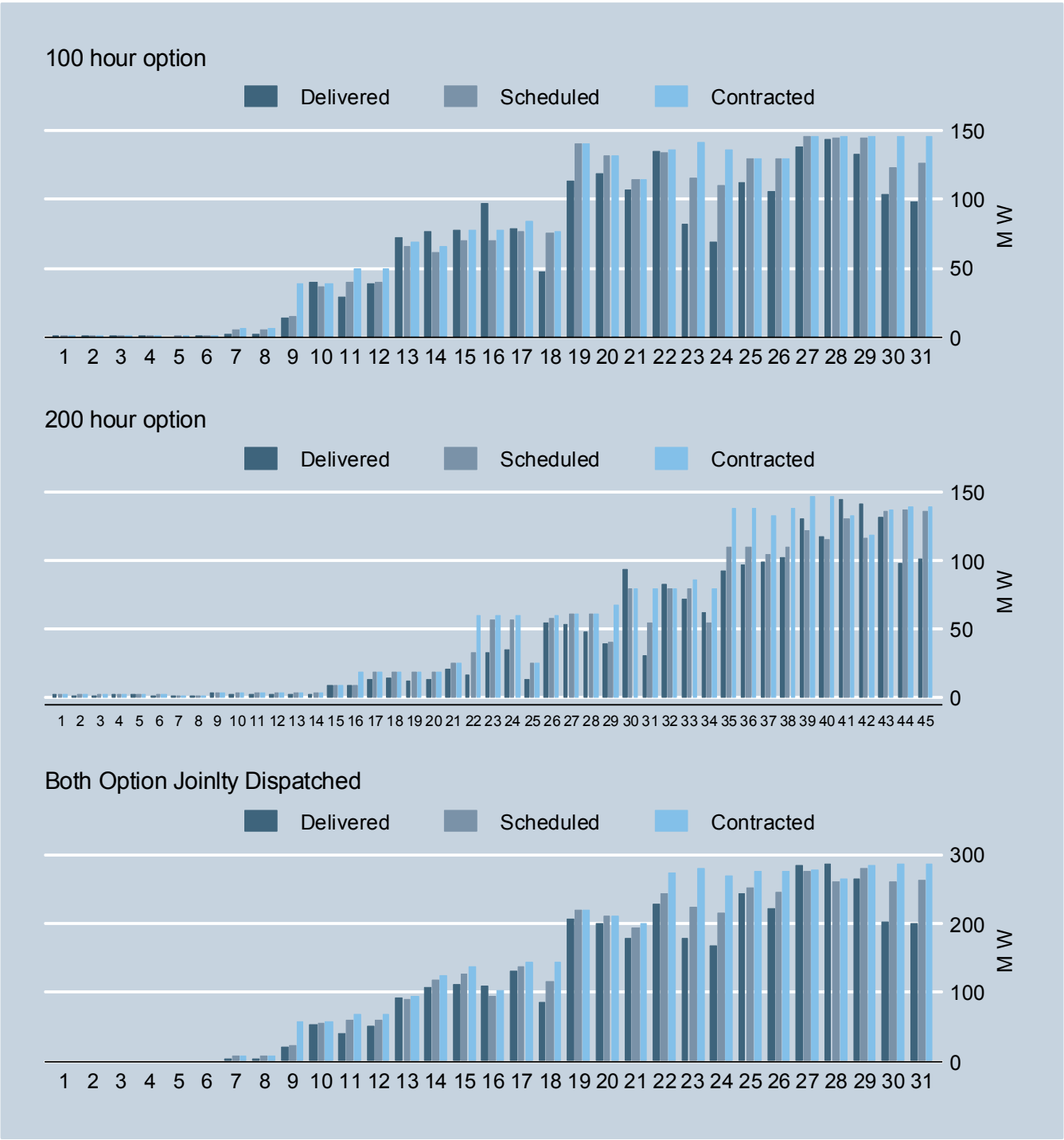

Figure 5: Event Performance based on Historical Events, by Dispatch Option (2008-2011)

Table 2 summarizes aggregator performance by year for each of the options. It contains the data underlying Figure 5. It is useful for assessing if aggregators improved at delivering the scheduled demand reductions or improved in making the contracted resources available. One would expect performance to improve and volatility to decrease as aggregators gain experience with their individual sites and as the resource grows and diversifies.

In general, 2008 performance is relatively high but includes so few sites that conclusions cannot be drawn based on that year. Aggregator resources performed better in 2010-2012 than they did in earlier years; a larger share of scheduled resources were delivered and larger share of contracted resources were available for operations. The influence of the program growth on event-to-event volatility is more subtle. The volatility is measured by the standard deviation of the percent of scheduled resources delivered and scheduled in each individual event day. While the level of performance increased, there was no clear indication of a change in volatility as the program expanded. However, the number of participants alone is a poor metric for diversity because the demand reduction resources remained highly concentrated among the largest customers.

Table 2: Aggregator Resource Performance by Year (2008-2012)

Option	Year	Events	Sites	Avg. Delivered	Avg. Scheduled	Avg. Contracted	% Delivered of Scheduled	% Scheduled of Contracted	Event Volatility in % Delivered	Event Volatility in % Scheduled
	2008	8	5.4	1.4	2.7	3.0	50.1%	92.2%	21.2%	6.2%
	2009	4	22.5	30.6	33.5	44.7	91.4%	74.9%	14.2%	23.5%
100 hour	2010	6	88.0	74.8	70.0	75.1	106.8%	93.3%	25.6%	3.4%
	2011	8	188.0	105.4	125.6	132.5	83.9%	94.8%	12.1%	8.7%
	2012	5	221.0	123.3	137.0	146.0	90.0%	93.8%	8.8%	7.6%
	2008	14	2.7	1.4	1.7	1.7	82.0%	100.0%	40.9%	0.0%
	2009	6	11.3	11.5	15.5	17.2	74.3%	90.3%	13.8%	21.8%
200 hour	2010	9	32.8	34.3	46.3	53.3	74.2%	86.9%	19.1%	18.3%
	2011	11	79.2	88.9	92.5	113.1	96.1%	81.8%	17.2%	10.8%
	2012	5	52.2	123.6	131.4	133.6	94.1%	98.4%	22.2%	0.4%
	2008	8	7.9	2.7	4.4	4.6	61.2%	95.0%	21.4%	4.6%
	2009	4	34.0	42.2	49.9	63.6	84.7%	78.4%	11.2%	24.1%
Both jointly dispatched	2010	6	118.3	106.8	114.3	124.9	93.4%	91.6%	13.3%	5.6%
	2011	8	275.0	203.4	225.5	250.8	90.2%	89.9%	7.1%	8.0%
	2012	5	273.2	246.9	268.3	279.5	92.0%	96.0%	15.0%	4.0%

A substantial share of the gap between delivered and scheduled resources can be attributed to systematic bias in settlement baselines. Using the out of sample process described in the methodology to assess accuracy, OPA evaluations have consistently identified a systematic bias of 6 to 7 percent in the settlement baseline.[3] While this bias may seem small, it translates into larger biases in the demand reductions delivered. To illustrate, a baseline that overestimates by 6 percent will estimate demand reduction of 36 percent when the actual demand reduction is 30 percent. In other words, although baseline error is 6 percent it overstates demand reductions by 20 percent (6%/30%).[4] The settlement

[3] The aggregate upward bias does not imply that a baseline over-estimates impacts for all aggregator settlement accounts. In fact, baseline errors tend to be larger for individual customers and for settlement accounts that are not diversified. Though a baseline is upwardly biased in aggregate, it still systematically underestimates demand reductions for a substantial share of settlement accounts; though in aggregate over-estimates outweigh under-estimates. Effective settlement rules minimize the total payment error, regardless of direction.

[4] This concept also applies to baselines that are downwardly biased. A baseline that is biased downward by -6% will estimate a reduction of 24% when the actual demand reduction is 30%.

rules for OPA's aggregator program are determined by averaging the event period for 15 days out of the prior 20 eligible days with the highest loads (15-in-20 baseline). Baseline rules are often determined prior to enrollment of customers and may need to be calibrated once that participant mix is better understood.[5] OPA adopted a 15-in-20 baseline in part because it anticipated enrollment of some weather sensitive customers, but in practice most of the demand reduction are delivered by extremely large industrial customers. Table 3 compares how aggregators performed as measured by the settlement baseline against how they performed as measured by OPA's annual evaluations.

Table 3: Comparison of Performance - Settlement versus Evaluation Impacts (2008-2012)

					Evaluation Impacts			Settlement Baseline		
Option	**Year**	**Events**	**Sites**	**Avg. Scheduled**	**Avg. Delivered**	**% Delivered**	**Performance Volatility**	**Avg. Delivered**	**% Delivered**	**Performance Volatility**
100 hour	2008	8	5.4	2.7	1.4	50.1%	21.2%	3.8	139.0%	57.9%
	2009	4	22.5	33.5	30.6	91.4%	14.2%	42.3	126.2%	14.2%
	2010	6	88.0	70.0	74.8	106.8%	25.6%	92.5	132.1%	11.5%
	2011	8	188.0	125.6	105.4	83.9%	12.1%	135.1	107.5%	12.4%
	2012	5	221.0	137.0	123.3	90.0%	8.8%	101.1	73.8%	12.0%
200 hour	2008	14	2.7	1.7	1.4	82.0%	40.9%	2.4	136.1%	55.6%
	2009	6	11.3	15.5	11.5	74.3%	13.8%	11.9	76.6%	14.6%
	2010	9	32.8	46.3	34.3	74.2%	19.1%	48.7	105.3%	30.0%
	2011	11	79.2	92.5	88.9	96.1%	17.2%	132.8	143.6%	38.1%
	2012	5	52.2	131.4	123.6	94.1%	22.2%	153.3	116.7%	21.8%
Both jointly dispatched	2008	8	7.9	4.4	2.7	61.2%	21.4%	6.2	141.6%	28.8%
	2009	4	34.0	49.9	42.2	84.7%	11.2%	54.3	108.9%	6.5%
	2010	6	118.3	114.3	106.8	93.4%	13.3%	140.9	123.2%	14.6%
	2011	8	275.0	225.5	203.4	90.2%	7.1%	278.4	123.4%	11.5%
	2012	5	273.2	268.3	246.9	92.0%	15.0%	254.5	94.8%	15.3%

Many aggregators manage to settlement baselines because they directly affect payments. That is, they calculate baselines in advance and reduce enough demand to comply with them. We raise this issue for two reasons. First, it is critical to measure baseline bias and understand how impacts estimated by baseline differ from actual demand reductions. Ideally, this is done prior to engaging in multi-year contracts. Second, it is useful to assess how aggregator resources performed according to the mechanism used to determine their payments. If aggregator resources perform well as measured by the settlement baseline, a substantial amount of underperformance can be eliminated through implementing a more accurate settlement rules.

Between 2009-2012, after the program matured, performance based on annual evaluations results—which explicitly tested multiple methods including settlement baselines and selected the most accurate one—range between 84.7 percent and 93.4 percent when both options are dispatched jointly. In contrast, when performance is measured according to settlement baselines impact estimates, aggregators reduce between 94.8 percent and 123.4 percent of the scheduled demand reductions. While the difference is due to the baseline bias discussed earlier, the key point is that aggregators generally over

[5] OPA had the additional disadvantage of not being able to access interval data unless customers are enrolled in one of its programs. In other words, it did not have the ability to assess how different baseline performed for different types of customers in advance.

perform relative to settlement baselines. Most aggregators will ensure enough demand reductions are delivered to avoid reduction in payments and, in fact, often hedge by over delivering according to the settlement baselines.

It is also useful to understand how performance of aggregator programs varies by factors such as industry, customer size and the amount of event experience at individual sites. Table 4 summarizes estimated performance by industry type. Table 5 summarizes estimated performance based on customer size. Both tables include performance based on the evaluation results and performance as measured by the settlement baselines. It is important to remember that aggregator obligations are by settlement accounts and not for individual customers. Put differently, the expected contribution for each customer is an estimate provided by the aggregators, not an obligation. The performance estimates rely on information aggregators provided to OPA regarding expected demand reductions for each site enrolled. These estimates were provided at the time each site was added to an aggregator's portfolio.

There are three industries for which results differ substantially: Auto Parts & Assembly, Industrial Tools/Metal Work/Electronics, and Other. The auto parts industry performance is generally lower, but the baseline results are also roughly 50 percent higher than the evaluation results. The low performance may be due to the turbulence in the industry between 2008 and 2012. The annual impact evaluations estimate lower performance by Industrial Tools/Metal Work/Electronics than the settlement baselines indicate. This is likely due to the fact that this industry is volatile, and the processes vary from day to day. The 15-in-20 baseline omits days when processes are shut down, while regressions estimate the likelihood that those processes are on. Finally, both the evaluation and baselines indicate that performance by sites classified as Other is low, though the estimated underperformance differs.

Table 4: Comparison of Performance by Industry (2008-2012)

Industry	Sites	Avg. Scheduled MW	Evaluation Results		Settlement Baseline	
			Delivered	Performance (%)	Delivered	Performance (%)
Agribusiness/Consumer Products/Plastics	41	12.4	8.5	68%	8.3	67%
Auto Parts & Assembly	37	18.8	9.6	51%	14.9	79%
Chemicals & Minerals/Gases & Liquids	26	34.2	36.9	108%	43.9	128%
Construction & Materials	28	28.2	36.8	130%	44.7	158%
Industrial Tools/Metal Work/Electronics	34	79.8	66.4	83%	103.8	130%
Offices, Hotels, Banks, Professional Services	153	13.8	14.6	106%	14.4	105%
Paper Products/Packaging/Textiles	17	15.7	14.5	92%	18.1	115%
Timber, Paper & Pulp	8	19.8	18.0	91%	16.1	82%
Wholesale & Transportation, Water Treatment	25	10.7	10.0	93%	12.4	116%
Other	58	24.4	17.0	70%	9.5	39%
Total	**427**	**257.8**	**232.2**	**90%**	**286.2**	**111%**

Table 5 reflects the extent to which the contractual obligations are concentrated among larger customers. Customers with average demand over 1 MW account for 85 percent of the contractual obligations even though they account for about a quarter of aggregator participants. Interestingly, the smallest customers and the largest customers performed the best. The table is also useful for assessing discrepancies between the baseline and evaluation impact estimates. For customers over 5 MW, the baseline results are 40 percent higher than the evaluation results. This is not surprising since large customers include customers that face wholesale electricity prices and have an incentive to reduce demand during hotter days when prices are higher. The regressions incorporate wholesale market prices into the reference loads; while the baselines do not.

Table 5: Comparison of Performance by Size (2008-2012)

Size Category (Based on historical average kW)	Sites	Avg. Scheduled MW	Evaluation Results		Settlement Baseline	
			Delivered	Performance (%)	Delivered	Performance (%)
200 kW or less	50	3.1	3.2	103%	5.3	171%
200 to 500 kW	122	12.0	10.4	87%	8.7	72%
500 kW to 1 MW	143	22.2	18.9	85%	7.2	33%
1 to 5 MW	81	77.0	54.7	71%	62.8	82%
5 MW or more	31	143.6	144.9	101%	202.2	141%
Total	**427**	**257.8**	**232.2**	**90%**	**286.2**	**111%**

Table 6 includes an assessment of performance based on the amount of event-day experience at individual sites. Older sites that have experienced more events generally outperform newer sites. A plausible explanation is that newer sites are still learning how to comply with calls to reduce demand while older sites have already gained this experience. Under this theory, one would expect newer sites to perform more reliably as they gain experience. However, because the analysis is based on observational data it is inadequate to conclude that additional experience leads to improved performance. Another competing explanation is customer self-selection: it is quite plausible that sites able to easily and predictably reduce demand were the first to enroll with aggregators.

Table 6: Comparison of Performance by Event Experience (2008-2012)

Number of Times Dispatched	Sites	Avg. Scheduled MW	Evaluation Results		Settlement Baseline	
			Delivered	Performance (%)	Delivered	Performance (%)
1 to 10 events	212	39.8	30.8	77%	23.4	59%
11 to 20 events	168	125.8	117.5	93%	154.0	122%
20 events or more	47	92.3	83.9	91%	108.8	118%
Total	**427**	**257.8**	**232.2**	**90%**	**286.2**	**111%**

Conclusions

The multi-year analysis of aggregator performance presented in this paper has several implications. It fills a gap created by the limited publicly available data on the reliability and performance of aggregator programs and includes a large enough number of events that reliable inferences can be drawn. There are five main implications from the analysis.

1) *Aggregators generally comply and often over perform with settlement rules*. Most aggregators will ensure enough demand reductions are delivered to avoid reduction in payments and, in fact, often hedge by over delivering according to the settlement baselines.
2) *Biased or inaccurate baselines create a dichotomy between actual and stated performance.* Baselines will not accurately measure performance if they are biased. In the case of OPA, although aggregators complied with baseline settlement rules, the reductions delivered were less than the contracted obligations. Utilities should conduct an analysis of baseline accuracy prior to entering into mid- or long-term contracts with aggregators. Because baseline rules are often

determined prior to enrollment of customers, they should include a clause allowing an update of the baselines method based an empirical analysis of accuracy.

3) *Aggregator resource performance generally improved with time.* A larger share of scheduled resources was delivered and a larger share of contracted resources was available for operations in later years.
4) *Performance can vary by industry, size and experience*. This finding is not surprising, but useful for understanding how to target aggregator resources.
5) *Diversity in the mix of participants is important*. In assessing the diversity of aggregator resources, analysts need to take into account both the number of sites and the concentration of loads.

Aggregator demand response resources have grown substantially and the reliability of their performance affects both grid operations and planning. We encourage other jurisdictions to explicitly analyze the accuracy of baselines, conduct multi-year analysis of performance, and publicly disclose results.

References

Bode, J., Malaspina, P., Hartman, L., Berghman, D. (2012) *2011 Impact Evaluation of Ontario Power Authority's Commercial & Industrial Demand Response Programs.* Prepared for: Ontario Power Authority. Available at: http://www.powerauthority.on.ca/evaluation-measurement-and-verification/evaluation-reports

Bode, J., Schellenberg, J., Malaspina, P., Hartman, L. (2011) *2010 Impact Evaluation of Ontario Power Authority's Commercial & Industrial Demand Response Programs.* Prepared for: Ontario Power Authority. Available at: http://www.powerauthority.on.ca/evaluation-measurement-and-verification/evaluation-reports

Bode, J., Schellenberg, J., Mangasarian, P. (2010) *2009 Impact Evaluation of Ontario Power Authority's Commercial & Industrial Demand Response Programs.* Prepared for: Ontario Power Authority. Available at: http://www.powerauthority.on.ca/evaluation-measurement-and-verification/evaluation-reports

Bode, J., Schellenberg, J., Mangasarian, P. (2009) *2007-2008 Impact Evaluation for Ontario Power Authority's DR-1 and DR-3 Programs.* Prepared for: Ontario Power Authority. Available at: http://www.powerauthority.on.ca/evaluation-measurement-and-verification/evaluation-reports

Braithwait, S.D., Hansen, D.G., Armstrong, D.A. (2013). *2012 Statewide Load Impact Evaluation of California Aggregator Demand Response Programs.* CALMAC Study ID: PGE0318. Available at www.calmac.org

Braithwait, S.D., Hansen, D.G., Armstrong, D.A. (2011). *2010 Statewide Load Impact Evaluation of California Aggregator Demand Response Programs.* CALMAC Study ID: PGE0318. Available at www.calmac.org

Braithwait, S.D., Hansen, D.G., Armstrong, D.A. (2010). *2009 Statewide Load Impact Evaluation of California Aggregator Demand Response Programs.* CALMAC Study ID: PGE0318. Available at www.calmac.org

FERC (2012). *2012 Assessment of Demand Response and Advanced Metering*. Available at: http://www.ferc.gov/legal/staff-reports/12-20-12-demand-response.pdf

George, S., Bode, J., Burwen, J. (2012) *2011 Statewide Evaluation of California Aggregator Demand Response Programs.* CALMAC Study ID: PGE0314. Available at www.calmac.org

PJM Interconnection (2012). *Emergency Demand Response (Load Management) Performance Report.* Available at: http://www.pjm.com/~/media/markets-ops/dsr/emergency-dr-load-management-performance-report-2012-2013.ashx