2022

Repeatability and Reproducibility Assessment of Residential Heat Pump Performance Evaluation Methodologies based on CSA EXP07 and AHRI 210/240

Parveen Dhillon

Xudong Wang

W. Travis Horton

James E. Braun

Repeatability and Reproducibility Assessment of Residential Heat Pump Performance Evaluation Methodologies based on CSA EXP07 and AHRI 210/240

Parveen DHILLON[1*], Xudong WANG[2], W. Travis HORTON[1], James E. BRAUN[1]

[1]Ray W. Herrick Laboratories, School of Mechanical Engineering, Purdue University
West Lafayette, 47907-2099, USA
pdhillon@purdue.edu; wthorton@purdue.edu; jbraun@purdue.edu

[2]Air Conditioning, Heating, and Refrigeration Institute, Arlington, VA 22201, USA
XWang@ahrinet.org

*Corresponding Author

ABSTRACT

In the current testing and rating approach for performance evaluation of residential heat pumps (e.g., AHRI 210/240), seasonal energy efficiency (e.g., SEER) is estimated based on equipment performance measured in steady-state tests at different conditions and a degradation factor to account for the cycling losses during part-load conditions. In this current methodology, equipment controls are overridden with proprietary control settings from the manufacturer. Even though this current methodology provides a metric to compare the relative performance of equipment from different manufacturers, it does not capture the interaction of the integrated controls, equipment, and building dynamics. To address this aspect, a load-based testing methodology (CSA EXP07) has been developed with the motivation of capturing realistic equipment performance in a laboratory setting while operating similar to field application conditions by allowing the equipment to respond to a simulated virtual building load. To fully examine the applicability of the developed methodology as next-generation rating standards, it is crucial to assess the repeatability and reproducibility of the load-based testing approach. In this paper, repeatability and reproducibility assessments of both test methodologies, CSA EXP07 and AHRI 210/240, are presented based on round-robin test results of five different heat pumps of varied sizes and types from two different test facilities. Here, a summary of the overall seasonal performance repeatability and reproducibility is provided for both test methodologies along with a root cause analysis for the observed differences and recommendations for a next-generation rating approach. Reasonable to good repeatability was observed in EXP07 results in both labs when only comparing repeated tests that didn't involve test unit re-installation in between. However, poor reproducibility was observed when units were tested in different labs. On the other hand, AHRI 210/240 test results showed overall good repeatability as well as reproducibility. Some of the important factors that may have led to poor reproducibility of EXP07 results include issues related to the installation and setup of the unit for testing (e.g., instrumentation, refrigerant charge, duct static pressure, etc.), different interpretations, and implementations of the EXP07 draft standard, and different characteristics of the environmental chambers. Although these factors are also relevant for the current standard (AHRI 210/240), their impact on the results for load-based testing might be more significant because the draft standard is new and more complicated and the dynamic behavior of the unit with its integrated controls may be sensitive to some installation effects. Some of the differences in implementation will likely be resolved as the standard matures and personnel becomes more familiar with its application. Some of the issues with differences in facilities could be addressed by facility and test equipment improvements.

1. INTRODUCTION

The current testing and rating procedure for residential direct expansion (DX) vapor compression systems in the US is based on AHRI 210/240 (AHRI, 2020) along with the method of test (MoT) outlined in ASHRAE Standards 37 and 116 (ASHRAE, 2010, 2019). The heat pump performance is measured in a pair of psychrometric chambers by keeping both the test rooms, indoor and outdoor, at a steady state for defined test conditions. In addition, compressor speeds and indoor unit airflow rates are also fixed with proprietary control settings from the manufacturer. As the current rating approach does not consider the embedded controls and their dynamic interaction with representative building

loads, it might not be representative of the test unit's actual field performance. To address some of the limitations of the current rating approach, a load-based testing methodology has been developed.

Hjortland and Braun (2019), Patil *et al.* (2018), Cheng *et al.* (2021) and Dhillon *et al.* (2022a) outline development, implementation, and demonstration of load-based testing that forms the basis for CSA (Canadian Standards Association) standard draft EXP07 (CSA, 2019) for residential equipment. In this methodology, the test unit overall dynamic performance is measured with its embedded controls and thermostat allowing it to respond to an emulated building load in the test lab that is representative of residential building applications. A sensitivity study of virtual building parameters and thermostat location on load-based test results was presented by Cheng *et al.* (2018). Dhillon *et al.* (2022c, 2018, 2021a) further implemented this approach to evaluate the performance of residential heat pumps and compared load-based test results to the AHRI 210/240 rating methodology. To validate how well the load-based testing approach characterizes heat pump performance in a lab compared to an actual residential building, Dhillon *et al.* (2021d, 2022e) compared the cooling and heating mode test results of a heat pump tested in a 2-story residential house to that of a laboratory using the load-based testing approach. Cremaschi and Perez Paez (2017) performed a feasibility study on a load-based testing methodology for unitary equipment with integrated economizers. Dhillon *et al.* (2021b) also proposed an alternative load-based testing methodology for RTUs with integrated economizers based on the virtual building model approach. The load-based testing approach can also be utilized to evaluate advanced heat pump control design in a test laboratory setting as demonstrated by Dhillon *et al.* (2021c) and Ma *et al.* (2021).

As discussed above, there have been several studies describing the development, implementation, and validation of load-based testing for residential air conditioners and heat pumps. However, to completely examine the developed methodology for next-generation rating standards, it is critical to assess its repeatability and reproducibility. Repeatability refers to the ability to obtain consistent performance results for a given test unit installed at a single facility with repeated tests, whereas reproducibility refers to obtaining consistent results across different facilities for the same test unit. Dhillon *et al.* (2022d) presented repeatability assessment of the load-based testing approach for cooling mode performance measurement of a 3-ton variable-speed heat pump based on the repeated tests in the same lab. Overall good repeatability was observed when comparing cooling seasonal efficiency based on test results; however, the study only considered a small set of data. The motivation of the study presented in this paper was to perform a comprehensive repeatability and reproducibility assessment of the load-based testing approach as per CSA EXP07 in cooling as well as heating modes based on the test results of multiple systems. Round-robin tests were conducted with multiple heat pumps of varied sizes and types in two test labs, UL and PG&E. To compare EXP07 results with the current standard testing and rating approach, round-robin tests were also conducted with some of these heat pumps based on the AHRI 210/240 testing procedure. In the sections below, first, an overview of the test units and the testing progression in two labs for both test methodologies is provided. Then, an overview of the data analysis approach is presented followed by the repeatability and reproducibility assessment of EXP07 and AHRI 210/240 based on different heat pump test results from two labs. Finally, the conclusion summarizes and critiques the findings, as well as makes recommendations for future studies.

2. METHODOLOGY

For repeatability and reproducibility analysis of CSA EXP07:19 (CSA, 2019) and AHRI 210/240-2023 (AHRI, 2020), five different units were tested multiple times at the same lab (UL), and then two of those units were tested at a different lab (PG&E) and the test results were compared within the same lab and in between different labs for each unit. Table 1 provides a brief description of the test units and their rated performance as per AHRI 210/240 (AHRI, 2017) including seasonal energy efficiency ratio (SEER) and heating seasonal performance factor (HSPF).

Table 1. Test units' description and rated performance

Test Unit	AFS	NEEA4	NEEA7 & NEEA7B	NRCan8	NRCan10
Description	5-ton split-system ducted (single-stage)	1-ton min-split ductless (variable-speed)	3-ton split-system ducted (variable-speed)	1.5-ton min-split ductless (variable-speed)	2-ton split-system ducted (variable-speed)
SEER	16	20	17.8	21.6	19
HSPF (Region IV)	8.5	12	11	11.7	10.5

Figure 1 shows the progression of testing as per EXP07 and AHRI 210/240 for different test units in the two labs. Each box named either CSA EXP07 or 210/240-2023 represents one complete set of cooling and heating test intervals based on the corresponding test methodology in the test lab highlighted on the left-hand side. In this paper "test interval" refers to a single test at a particular test condition and "test" refers to the entire set of test intervals as per the test methodology associated with either EXP07 or AHRI 210/240. For example, as per EXP07, one test consists of 5 cooling dry coil, 4 cooling humid coil, 6 heating continental, and 4 heating marine test intervals (CSA, 2019). In Figure 1, the number in the box on the right side adjacent to each test shows the corresponding test name or number. This test name or number in combination with the test unit and test lab provides a unique identifier for each set of test intervals data. Note that while referring to test names with "PREV" in their name, a short form "P" is generally used in the text in place of "PREV", for example, "PREV1" and "P1" refer to the same test. Figure 1 also highlights whether a test unit was taken out of test rooms and re-installed in between different tests either in the same or different lab. One thing to note regarding NEEA7 is that after the initial phase of UL testing when it was shipped to PG&E, it was discovered during setup that the unit was not operating properly. As a result, a different unit but the same model, NEEA7B, was used for PG&E tests and one final test at UL. Even though the test unit was changed from NEEA7 to NEEA7B in the test sequence, the results of these two different units were still compared for overall repeatability and reproducibility assessment of EXP07 and AHRI 210/240 as the same model was utilized in testing. Further, here NEEA7(B) is used to indicate that the analysis results include the test results of both test units NEEA7 and NEEA7B.

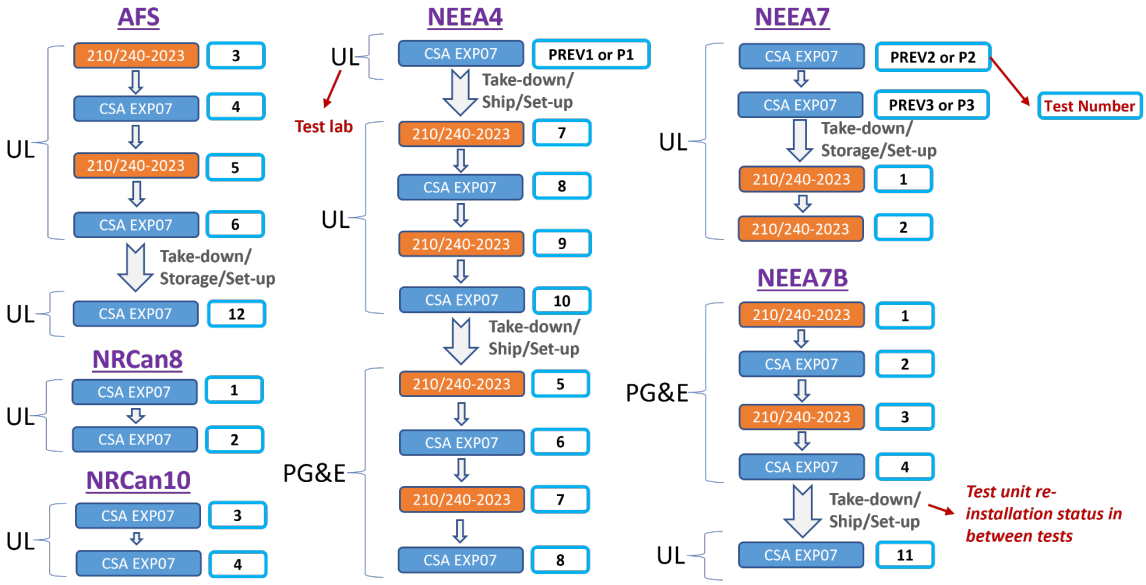

Figure 1. Round robin tests sequence for repeatability and reproducibility analysis of CSA EXP07:19 and AHRI 210/240-2023

Table 2. Dataset nomenclature approach

Test Unit	–	Test Lab	–	Test Methodology	–	No. of Tests in Dataset	|	Unit Reinstallation Status between Tests
AFS **NRCan8** **NRCan10** **NEEA4** **NEEA7** **NEEA7B** **NEEA7(B):** NEEA7 & NEEA7B		**UL** **PGE:** PG&E **UL&PGE:** UL and PG&E		**E:** EXP07-2019 **A:** AHRI 210/240-2023		**2** **3** **4** **5** (Individual tests included in each dataset are provided in Table 3)		**R:** Test unit reinstalled at least once in between different tests in the dataset, either in the same lab or in different labs **W:** Tests were performed without test unit reinstallation in between different tests in the dataset

For CSA EXP07 and AHRI 210/240 repeatability assessment, each test unit's performance across different tests from the same lab is compared, whereas, for reproducibility assessment, NEEA4 and NEEA7(B) (NEEA7 & NEEA7B)

performance from different tests across two labs are compared. For repeatability and reproducibility evaluation, different datasets are defined based on the nomenclature provided in Table 2. Table 3 shows the defined datasets for EXP07 and AHRI 210/240 repeatability and reproducibility evaluation. For EXP07 repeatability assessment, there are 9 datasets, 7 based on tests done at UL and 2 based on tests done at PG&E. It should be noted that with AFS and NEEA4, three tests were performed based on EXP07 at UL and the test unit was once removed and reinstalled in test rooms in between those three tests. Similarly, with NEEA7, the first two tests (P2 and P3) were performed based on EXP07 at UL and then one final test #11 with NEEA7B was performed at UL. So, the repeatability evaluation results using all UL tests for AFS, NEEA4, and NEEA7(B) are somewhat hybrid between true repeatability and reproducibility results as the unit was once removed and re-installed in between repeated tests. As can be seen in Table 3, two different datasets are defined for UL EXP07 testing of both units NEEA4 and NEEA7(B) in order to assess the impact of reinstallation on repeatability. In contrast, the EXP07 PG&E repeatability results for the same unit models are based on tests that were repeated without test unit reinstallation. For AHRI 210/240 repeatability evaluation, there are 5 datasets across two labs. In addition, for EXP07 and AHRI 210/240 reproducibility assessment, there are two datasets for each using all the tests at UL and PG&E for NEEA4 and NEEA7(B). There are 5 tests in each dataset for EXP07 reproducibility assessment and 4 tests in each dataset for AHRI 210/240 reproducibility evaluation.

Table 3. Datasets for EXP07 and AHRI 210/240 repeatability and reproducibility analysis based on different sets of tests in two labs for multiple units

	Dataset	Test Unit	Test Lab [Tests in Dataset]	Test Methodology
Repeatability	AFS-UL-E-3R	AFS	UL [4 & 6 & 12]	EXP07:2019
	NRCan8-UL-E-2W	NRCan8	UL [1 & 2]	EXP07:2019
	NRCan10-UL-E-2W	NRCan10	UL [3 & 4]	EXP07:2019
	NEEA4-UL-E-3R	NEEA4	UL [P1 & 8 & 10]	EXP07:2019
	NEEA4-UL-E-2W	NEEA4	UL [8 & 10]	EXP07:2019
	NEEA4-PGE-E-2W	NEEA4	PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL-E-3R	NEEA7 & NEEA7B	UL [P2 & P3 & 11]	EXP07:2019
	NEEA7-UL-E-2W	NEEA7	UL [P2 & P3]	EXP07:2019
	NEEA7B-PGE-E-2W	NEEA7B	PG&E [2 & 4]	EXP07:2019
	AFS-UL-A-2W	AFS	UL [3 & 5]	AHRI 210/240-2023
	NEEA4-UL-A-2W	NEEA4	UL [7 & 9]	AHRI 210/240-2023
	NEEA4-PGE-A-2W	NEEA4	PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7-UL-A-2W	NEEA7	UL [1 & 2]	AHRI 210/240-2023
	NEEA7B-PGE-A-2W	NEEA7B	PG&E [1 & 3]	AHRI 210/240-2023
Reproducibility	NEEA4-UL&PGE-E-5R	NEEA4	UL [P1 & 8 & 10], PG&E [6 & 8]	EXP07:2019
	NEEA7(B)-UL&PGE-E-5R	NEEA7 & NEEA7B	UL [P2 & P3 & 11], PG&E [2 & 4]	EXP07:2019
	NEEA4-UL&PGE-A-4R	NEEA4	UL [7 & 9], PG&E [5 & 7]	AHRI 210/240-2023
	NEEA7(B)-UL&PGE-A-4R	NEEA7 & NEEA7B	UL [1 & 2], PG&E [1 & 3]	AHRI 210/240-2023

Table 4. EXP07 cooling and heating climate zones and test type used in SCOP estimation

Cooling Climate Zone	**Very Cold**	**Cold/Dry**	**Cold/ Humid**	**Marine**	**Mixed**	**Hot/ Humid**	**Hot/Dry**	
Test Type	Humid	Dry	Humid	Dry	Humid		Dry	
Heating Climate Zone	**Subarctic**	**Very Cold**	**Cold/Dry**	**Cold/ Humid**	**Marine**	**Mixed**	**Hot/ Humid**	**Hot/Dry**
Test Type	Continental				Marine	Continental		

As per EXP07, a test unit's performance is measured in two sets of cooling test intervals, dry coil and humid coil, and two sets of heating test intervals, continental and marine, which represent different climate types (CSA, 2019). Then,

measured performance is propagated through a temperature bin method to estimate the cooling and heating seasonal coefficient of performance (SCOP) for different climate zones. Table 4 shows different cooling and heating climate zones as per EXP07 along with corresponding test type results used in SCOP estimation. For AHRI 210/240 testing, a test unit performance is measured at one set of cooling and heating test conditions which are also propagated through a temperature bin method to estimate SEER and HSPF for a single climate zone (AHRI, 2020) rather than multiple climate zones as in EXP07.

2.1 Analysis Approach

For a test methodology repeatability and reproducibility evaluation, both individual test interval and seasonal performance metrics are compared to assess the overall performance variability. For a quantitative and qualitative comparison of performance in different datasets as per EXP07, measured COP and observed test unit dynamic behavior during convergence are compared for each test interval along with a root cause analysis of the observed performance differences. Overall estimated cooling and heating SCOPs were also compared between different tests across different climate zones. A similar data analysis approach was used to assess AHRI 210/240-2023 repeatability and reproducibility.

To quantify the variation in a performance metric, three different statistical parameters were utilized. One is the difference in the minimum and maximum value of a performance metric between different tests to illustrate the maximum variation. Second is the 95% confidence interval (CI) for the average performance metric estimated based on the results from different tests. A student's t-distribution is assumed for the test unit average performance metric to estimate the 95% confidence interval as there were a limited number of samples in each dataset. However, the t-distribution confidence interval is also a strong function of the number of samples, which might lead to a larger confidence interval with a smaller number of samples (tests) for a unit even though the overall variation in measured performance is relatively small. Thus, using confidence intervals to compare the performance variation results between two units with a different number of tests can lead to erroneous conclusions. Therefore, a third statistical parameter, population standard deviation (STD), was also used to quantify the performance variation in different tests.

3. REPEATABILITY AND REPRODUCIBILITY ASSESSMENT RESULTS

In the subsections below, first, EXP07 repeatability and reproducibility evaluation results are presented for different test units in two labs, along with root cause analysis for the observed differences and recommendations to further improve the load-based testing approach. Following that, AHRI 210/240 repeatability and reproducibility evaluation results are discussed. Then a comparison of the two test methodologies repeatability and reproducibility is presented. For more detailed results and analysis, the reader is referred to the AHRI 8026 project report (Dhillon *et al.*, 2022b).

3.1 Load Based Testing Approach (EXP07)

Figure 2 presents repeatability assessment based on 7 different datasets (Table 3) for various test units tested at UL showing the minimum, maximum, and mean of two statistical parameters across different climate zones (Table 4) for estimated cooling and heating SCOP based on repeated tests. ΔSCOP (Max-Min) is the difference in the minimum and maximum estimated SCOP of different tests for each climate zone described as the percentage of the tests' mean SCOP. It represents the maximum variation in performance between different tests. SCOP STD is the standard deviation of SCOP based on repeated tests shown as a percentage of the average SCOP of multiple tests. This represents the overall variation in estimated seasonal performance normalized to the number of repeated tests since the number of repeated tests varied in different datasets for different units as shown in Figure 1 and Table 3. Overall, good repeatability was observed in cooling and heating estimated SCOP of AFS, NRCan8, and NRCan10 test units at UL as shown with results based on corresponding datasets for these units. A relatively larger variation was observed in measured performance among three tests in dataset NEEA7(B)-UL-E-3R with the NEEA7 and NEEA7B unit, where reasonable repeatability was observed with somewhat better results in heating mode compared to cooling mode tests. Furthermore, good repeatability is achieved in both heating and cooling mode when considering the NEEA7-UL-E-2W dataset. That dataset only considers back-to-back tests (P2 and P3) at UL with NEEA7 without including the last test #11 with NEEA7B at UL that was performed after reinstallation about two years later.

NEEA4 test results for dataset NEEA4-UL-E-3R showed somewhat poor repeatability, which was mostly due to significantly different measured performance in test P1 which was performed more than a year earlier compared to the other two sets of tests, #8 and #10 and the test unit was taken out of the test rooms after test P1 and re-installed.

Without taking test P1 into account in repeatability evaluation in dataset NEEA4-UL-E-2W, the NEEA4 shows good repeatability in cooling mode and reasonable repeatability in heating mode. The difference in unit performance during test P1 compared to the other two tests is mostly attributable to differences in test unit dynamic response under the same test conditions and unit dynamic behavior captured during the convergence period in some test intervals. In addition to the variability in test unit integrated controls response, differences in NEEA4 dynamic response at the same test conditions could be due to differences in the test setup, charge, testing approach, etc. It should also be noted that tests P2 and P3 with NEEA7 were also performed around the same time as test P1 with NEEA4 and around 24 months before the 3rd test (#11) with NEEA7B. However, unit NEEA7(B) (NEEA7 & NEEA7B) based on dataset NEEA7(B)-UL-E-3R showed better repeatability than NEEA4 with dataset NEEA4-UL-E-3R.

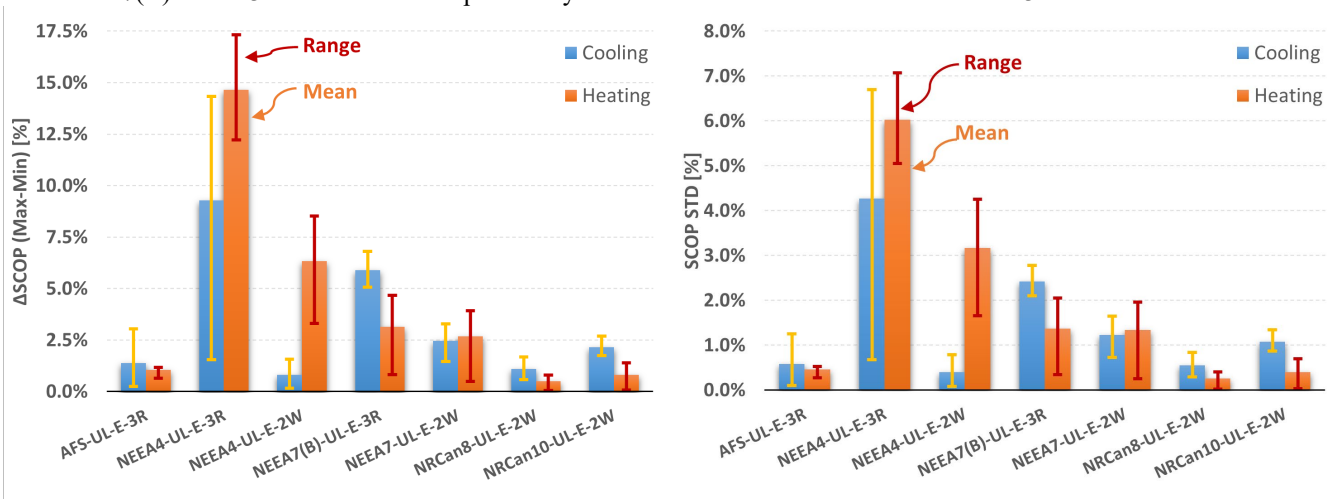

Figure 2. EXP07 cooling and heating SCOP repeatability assessment summary for different datasets from UL with mean and range of statistical parameters across different climate zones

Figure 3 shows EXP07 repeatability assessment summary results based on datasets of two test units at PG&E. Good overall repeatability was observed for both test units, NEEA4 and NEEA7B, in cooling as well as heating mode. In comparison to UL results, PG&E test results for both units showed better repeatability. This could be because, unlike at UL, the PG&E tests were performed back-to-back with no unit re-installation in between. As a result, there was no variability associated with the test setup, refrigerant charge, instrumentation, and other similar factors.

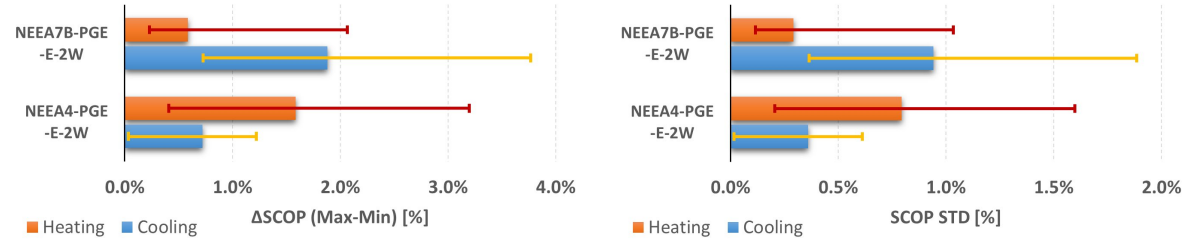

Figure 3. EXP07 cooling and heating SCOP repeatability assessment summary for two datasets from PG&E with mean and range of statistical parameters across different climate zones

Figure 4 shows summary results of the EXP07 reproducibility assessment for two test units, based on results from multiple tests at UL and PG&E. In NEEA4 cooling tests at PG&E, full-load tests were erroneously performed in some test intervals where load-based tests should have been conducted, due to a possible issue with the testing approach implementation. This resulted in some large differences in cooling SCOP between the two labs, and because of this testing approach issue, comparing NEEA4 cooling SCOP between the two labs does not provide relevant conclusions regarding EXP07 reproducibility assessment, and thus results for this case are not shown in the plots. Furthermore, heating test results for NEEA4 showed poor reproducibility. Compared to NEEA4, NEEA7(B) demonstrated better reproducibility in both cooling and heating results. Overall good reproducibility was observed for NEEA7(B) heating test results, whereas the test unit showed relatively poorer reproducibility for cooling test results. Table 5 provides a categorization of different factors that contribute to the overall repeatability and reproducibility performance variations of the EXP07 test results along with a description of their influences and variability.

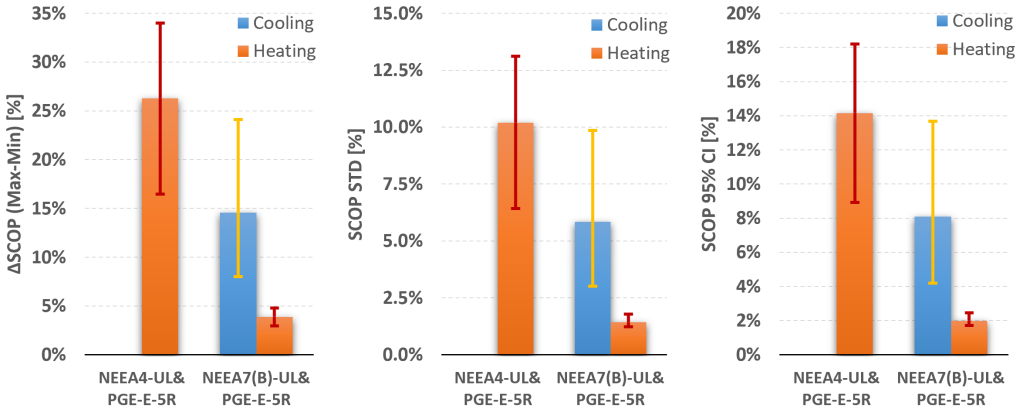

Figure 4. EXP07 cooling and heating SCOP reproducibility assessment summary for two test units for testing at UL and PG&E with mean and range of statistical parameters across different climate zones

Table 5. Categorization of variables contributing to EXP07 repeatability and reproducibility

No.	Factor	Repeatability Issues	Reproducibility Issues
1	**Unit under Test (UUT) Installation/Setup** - includes differences in refrigerant charge, duct static pressure setup, thermostat offset setup, type & installation of instrumentation	Relevant when unit is re-installed between tests, especially when installations occur many months apart, when different personnel are involved, and if interpretation of the standard setup changes	An inherent issue for any test standard, but load-based testing has additional installation requirements that are less familiar to personnel and that are still evolving
2	**Environmental Chamber Characteristics** - includes differences in air flow and temperature distribution, the responsiveness of reconditioning controls	Potentially an issue if a UUT is re-installed within a different chamber between tests or at a different location within the same chamber; also an issue if chamber controls do not track indoor or outdoor setpoints well due to UUT dynamic behavior	Differences in airflow and temperature distribution could impact the dynamic response of the thermostat; in some cases, chamber controls may not track indoor or outdoor setpoints well due to UUT dynamic behavior
3	**Interpretation and Implementation of Method of Test** - includes convergence criteria, transition to full-load tests	Especially relevant if different personnel are involved in the implementation and/or when repeatability tests are separated over a long duration. Also, there are some shortcomings with the current convergence criteria	An inherent issue for any test standard, but load-based testing has additional requirements that are less familiar to personnel and that are still evolving
4	**Dynamic Behavior of UUT** - covers variations in the dynamic behavior of UUT due to variations in control behavior that may result from adaptive learning and the limited time available for testing	Lack of repeatability due to inconsistent control behavior is an inherent issue for some equipment that probably needs to be addressed by each manufacturer	Lack of reproducibility due to inconsistent control behavior is an inherent issue for some equipment that probably needs to be addressed by each manufacturer
5	**UUT Replacement** - due to failure or performance degradation	Performance varies due to manufacturing tolerances and/or firmware differences	Performance varies due to manufacturing tolerances and/or firmware differences

Table 6 provides a qualitative evaluation of repeatability results for different datasets along with the possible factors as per Table 5 contributing to the observed differences between repeated tests. The factors considered to be the most important in contributing to observed large differences between different tests are described. As mentioned previously, the UL repeatability results for AFS, NEEA4, and NEEA7(B) based on all three repeated tests at UL, as shown with

datasets AFS-UL-E-3R, NEEA4-UL-E-3R, and NEEA7(B)-UL-E-3R are somewhat between true repeatability and reproducibility results because the units were re-installed once in between repeated tests. This could be one of the primary factors for relatively poor repeatability with the NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R datasets at UL compared to other datasets. For both units, when only considering two sets of back-to-back tests at UL, in datasets NEEA4-UL-E-2W and NEEA7-UL-E-2W, repeatability improves for both cooling and heating modes.

Table 6. EXP07 repeatability assessment summary for different datasets in two labs along with key factors contributing to observed differences

Dataset	Repeatability		Important Factors
	Cooling	Heating	
AFS-UL-E-3R	Good	Good	**1** - the unit was re-installed after being in storage for some time and then retested
NEEA4-UL-E-3R	Poor	Poor	**1, 3, 4** - unit was re-installed after some time and then re-tested; convergence criteria implementation; UUT had inconsistent dynamic response and also a difference in measured airflow at same test conditions
NEEA4-UL-E-2W	Good	Fair	**3, 4** - differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at the same test conditions
NEEA7(B)-UL-E-3R	Fair	Good	**1, 3, 4, 5** – unit re-installed after a while and re-tested; UUT was replaced due to some issues while setting up for testing at PG&E lab; UUT had an inconsistent dynamic response; also had differences in dynamic response captured with convergence criteria
NEEA7-UL-E-2W	Good	Good	**3, 4** - some differences in dynamic response captured with convergence criteria in different tests; UUT had inconsistent dynamic response at the same test conditions
NRCan8-UL-E-2W	Good	Good	-
NRCan10-UL-E-2W	Good	Good	-
NEEA4-PGE-E-2W	Good	Good	-
NEEA7B-PGE-E-2W	Good	Good	-

Table 7. EXP07 reproducibility assessment summary for two datasets along with key factors contributing to observed differences

Dataset	Reproducibility		Important Factors
	Cooling	Heating	
NEEA4-UL&PGE-E-5R	-	Poor	**1, 2, 3, 4** - differences associated with the facilities and UUT installation/setup including thermostat offset settings and measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also difference in the transition to full-load.
NEEA7(B)-UL&PGE-E-5R	Fair	Good	**1, 2, 3, 4, 5** - differences associated with the facilities and UUT installation/setup; measured airflow; UUT had inconsistent dynamic response; differences in convergence criteria implementation and dynamic response captured during convergence period; also, UUT was replaced in between tests.

It should be clear from the results of Table 6 and Table 7 that performance results are much less consistent after equipment has been re-installed in either the same or a different laboratory than if the equipment is retested with the same installation. Some of the important factors that affect the reproducibility include issues related to the installation and setup of the unit for testing (e.g., instrumentation, refrigerant charge, duct static pressure, etc.), different interpretations, and implementations of the EXP07 draft standard, and different characteristics of the environmental chambers. Although these factors are also relevant for the current standard (AHRI 210/240), their impact on the results for load-based testing might be more significant because the draft standard is new and more complicated and the

dynamic behavior of the unit with its integrated controls may be sensitive to some installation effects. In addition to the effect of installation and implementation issues on test unit dynamic behavior, variations in test unit dynamic response at the same test conditions could be due to the adaptive/learning behavior of the embedded controller and the limited time available for testing. Some of the differences in implementation will likely be resolved as the standard matures and personnel becomes more familiar with its application. Some of the issues with differences in facilities could be addressed by facility and test equipment improvements. For example, utilizing a thermostat apparatus (Kim *et al.*, 2022) to provide a standardized environment for the thermostat could reduce differences that are caused by non-uniform airflow and temperatures within environmental chambers. However, differences related to variations in test unit controller behavior would be challenging to address with the test standard without dramatically increasing the testing time. In fact, one of the merits of the load-based testing approach is that it can capture the impacts and sensitivities of test unit performance to dynamic responses of its embedded controller. If a test unit controller performs inconsistently with load-based testing in the laboratory, then it is likely to perform similarly in the field and it is important to capture these effects in a standard method of test. This could be an incentive for manufacturers to develop controllers with more consistent and predictable behavior. With that being said, it would be a good idea to further investigate how best to test and rate units that have inconsistent controller behavior using load-based testing in a repeatable and representative fashion for making the standard more inclusive.

There were likely some differences in the implementation of EXP07 convergence criteria, which defines the procedure to select the converged results to calculate the test unit performance for a test interval, which led to some significant performance differences between tests carried out in different labs. For instance, differences in test unit operation mode and dynamic behavior that were captured during periods that were deemed to be converged were likely due to different interpretations of the EXP07 convergence criteria and some limitations of the current convergence criteria. These convergence criteria issues occurred when there was a combination of different operating modes that occurred at a given test condition, such as a sequential combination of a unit on/off cycling and defrost, or when there was irregular on/off cycling pattern with a mixture of short and regular cycles. These issues are discussed in detail in the report (Dhillon *et al.*, 2022b) with recommended improvements that could be used as a reference to update the EXP07 convergence criteria. Another specific issue identified through testing was that indoor temperatures could converge to significantly different values, most notably when comparing test results between labs. This could primarily be related to differences in the thermostat offset settings at the start of testing that is part of the implementation of the EXP07 setup procedures. Another EXP07 implementation issue that led to disparities between laboratories is related to the interpretation of the decision of when to switch to full-load testing. Prematurely switching to full-load testing can have a significant impact on results at relatively high-load conditions that are important in determining seasonal efficiencies.

Some of the possible sources of differences in test unit performance related to test setup (e.g., charge, thermal mass, instrumentation location, sensor dynamic responses, air flow measurement, test unit fan settings) and test facility control (e.g., room conditions, external static pressure) could be addressed by improving test setup requirements and/or corrections for the components that have a significant effect on dynamic performance measurements, such as code-tester thermal mass, instrumentation location, sensor response, etc. Also, more appropriate test conditions and operating tolerances should be defined to limit the effect of variations in test facility control on overall performance measurement. Both of these enhancements, however, will necessitate additional research and examination because most of the currently used test methods are for steady-state tests rather than dynamic load-based tests. Also, issues with external static pressure control, airflow measurement, and test fan settings that led to differences in the two labs should be further investigated to update the current EXP07 testing approach to have better reproducibility.

3.2 Steady State Testing Approach (AHRI 210/240)

In this section, an overall summary of the AHRI 210/240 repeatability and reproducibility assessment is provided based on test results of different units in two labs. Figure 5 shows the average SEER and HSPF for each unit based on corresponding datasets as defined in Table 3 from UL and PG&E labs along with population standard deviation (STD) to show variability in estimated seasonal performance based on different tests. The AFS unit showed good repeatability at UL in two repeated tests as shown with the AFS-UL-A-2W dataset results. Good repeatability was also observed for the NEEA4 unit at UL and PG&E as shown with datasets NEEA4-UL-A-2W and NEEA4-PGE-A-2W, except for cooling test results at UL in dataset NEEA4-UL-A-2W. In cooling test intervals at UL with NEEA4, relatively poor performance was measured in test #7 compared to test #9 in dataset NEEA4-UL-A-2W. Further, for the NEEA4 unit, higher SEER was estimated based on PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, resulting in relatively poor reproducibility in the cooling test results based on dataset NEEA4-UL&PGE-A-4R. However, the NEEA4 unit had slightly better HSPF that was estimated based on

PG&E tests in dataset NEEA4-PGE-A-2W compared to UL tests in dataset NEEA4-UL-A-2W, and overall good reproducibility was noted in heating test results based on dataset NEEA4-UL&PGE-A-4R. NEEA7 and NEEA7B test results showed good repeatability in the UL tests in dataset NEEA7-UL-A-2W as well as in PG&E lab tests in dataset NEEA7B-PGE-A-2W. At UL, relatively better repeatability was noted in cooling test results compared to heating in dataset NEEA7-UL-A-2W, whereas the opposite was observed in PG&E test results in dataset NEEA7B-PGE-A-2W. Between the two labs, overall good reproducibility was observed for cooling as well as heating results based on dataset NEEA7(B)-UL&PGE-A-4W. The differences in the two labs are possibly due to differences in the test setup, charge, instrumentation, airflow rate measurement, external static pressure control, and measurement uncertainties.

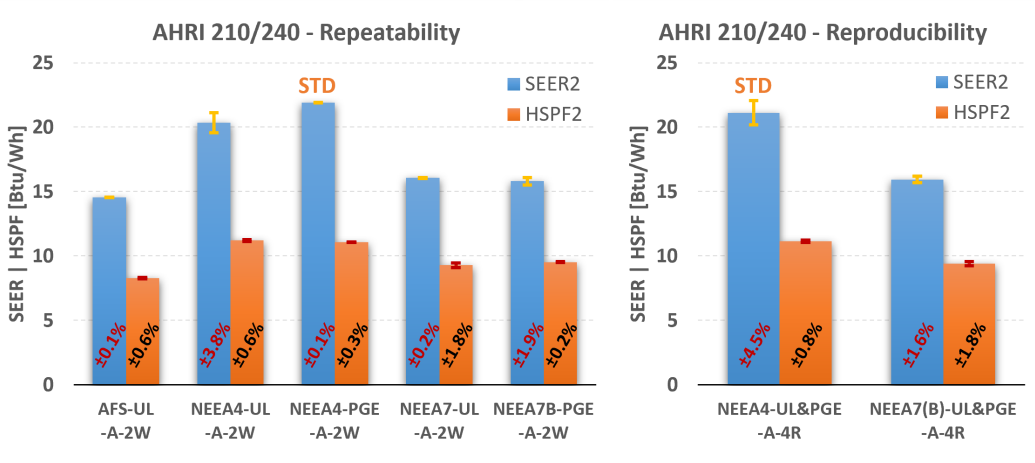

Figure 5. AHRI 210/240 repeatability and reproducibility summary with average seasonal performance metric average and STD based on different tests

3.3 Comparison

Figure 6 and Figure 7 compare the repeatability and reproducibility of EXP07 and AHRI 210/240 in terms of percentage STD applied to cooling and heating seasonal performance metrics using the test data from UL and PG&E for NEEA4 and NEEA7(B) (NEEA7 & NEEA7B). STD is used as a statistical parameter rather than 95% CI because a different number of tests were used in the AHRI 210/240 and EXP07 assessments for some cases. For EXP07 results, a mean and range of cooling and heating SCOP STD observed across different climate zones are shown, whereas single SEER and HSPF STD values are presented for AHRI results for a single climate zone. The results for reproducibility can be assessed by viewing the results for all tests of each unit across both labs in Figure 6 and Figure 7. The datasets utilized for EXP07 and AHRI 210/240 reproducibility assessment for NEEA4 and NEEA7(B) can be seen in Table 3. For both NEEA4 and NEEA7(B), significantly better reproducibility is observed in AHRI 210/240 results compared to EXP07, except for NEEA7(B) heating results, where comparable variations in seasonal performance metrics were observed for both testing approaches. Some possible reasons for the poor reproducibility of EXP07 have previously been discussed.

Repeatability can be assessed by viewing results in Figure 6 and Figure 7 that were determined from test results at each lab (UL and PG&E). Note that the AHRI 210/240 repeatability tests for units NEEA4 and NEEA7(B) involved only two tests at each lab that were performed sequentially without having to reinstall the unit as defined in datasets NEEA4-UL-A-2W, NEEA4-PGE-A-2W, NEEA7-UL-A-2W, and NEEA7B-PGE-A-2W. The EXP07 tests at UL involved three tests where only two of the tests were back-to-back without reinstallation as defined in datasets NEEA4-UL-E-3R and NEEA7(B)-UL-E-3R. To provide fairer repeatability comparisons, results for UL EXP07 that only include back-to-back tests without reinstallation are also shown (NEEA4-UL-E-2W and NEEA7-UL-E-2W datasets). Overall, the repeatability of both EXP07 and AHRI 210/240 are good and comparable in both cooling and heating when not considering data where the unit was not reinstalled. In fact, there are some examples where the EXP07 repeatability was better than for AHRI 210/240 (e.g., cooling tests for NEEA4 at UL). If data is included for units that were reinstalled at UL, then the repeatability of AHRI 210/240 would generally be better than that for EXP07, especially for the NEEA4 heating results. However, it is felt that these comparisons are less relevant for repeatability because they contain several factors related to reproducibility.

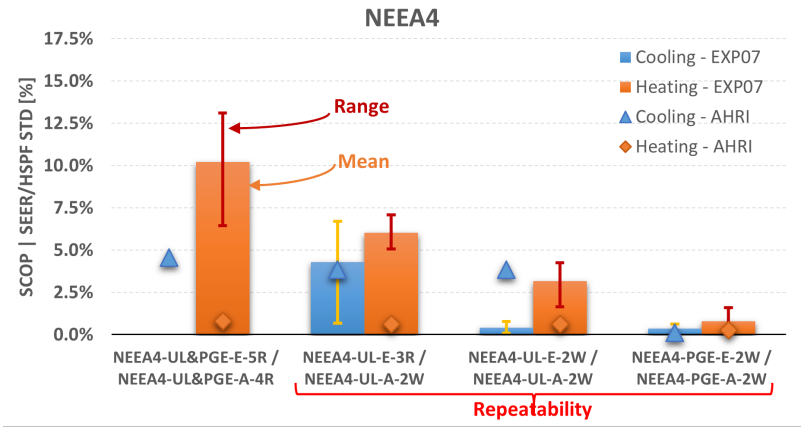

Figure 6. NEEA4 EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for a single climate zone

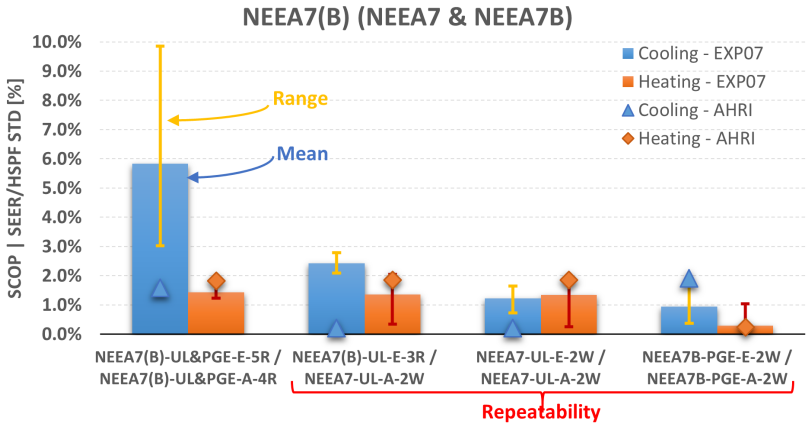

Figure 7. NEEA7(B) EXP07 and AHRI 210/240 seasonal performance metrics STD comparison for reproducibility and repeatability assessment. EXP07 SCOP STD mean and range across different climate zones and AHRI SEER / HSPF STD for a single climate zone

4. CONCLUSIONS

In this paper, repeatability and reproducibility assessments of two testing and rating methodologies for residential unitary equipment are presented. One is based on a newly developed load-based testing approach as per CSA EXP07:19, whereas the other is based on the current testing and rating approach as per AHRI 210/240-2023. Round robin test data of five different heat pumps tested across two different labs, UL and PG&E, were analyzed to perform a quantitative and qualitative comparison of measured performance in different tests at the same test conditions. For each test unit, performance measured in each test interval as well as overall estimated seasonal performance metrics were compared for repeatability evaluation in each lab and overall reproducibility across two labs for both test methodologies. Reasonable to good repeatability was observed in EXP07 results in both labs when only comparing repeated tests that did not involve test unit re-installation in between. However, poor reproducibility was observed except for the heating mode test results of NEEA7(B) unit. On the other hand, AHRI 210/240 test results for both units showed good repeatability as well as reproducibility. A root cause analysis of the observed performance variations was conducted, and possible causes for observed differences were highlighted. In addition, some areas of improvement for the EXP07 load-based testing methodology were also identified. These are mostly related to convergence criteria, test unit setup, transition to full-load testing, unit control settings, airflow measurement, thermostat offsets, and tolerances associated with dynamic measurements. Many of the improvements may have already been addressed in the most recent updates to EXP07:19 that have led to the latest version (EXP07:22) that is scheduled to be published in August 2022. It is strongly recommended that a similar study be carried out to assess the overall impact of updates that are incorporated in EXP07:22.

A key question that needs to be addressed in future work is: what is an acceptable tolerance in seasonal performance metric repeatability and reproducibility for load-based testing per EXP07? An initial step that is necessary to answer this question involves carrying out a detailed uncertainty analysis for measured performance that is due to dynamic measurements that are subject to both instrumentation uncertainties and dynamic behavior associated with the dynamic testing approach, the testing facilities, and operating tolerances. This should lead to the definition of minimum tolerances. However, the actual tolerances would need to be larger to account for human factors that lead to differences in test installation and setup in different labs. These tolerances can only be defined through testing data obtained across multiple labs. It is also important to analyze AHRI 210/240 and EXP07 test results to identify the primary sources of differences between the two approaches along with their contributions to overall estimated seasonal performance differences. It is also crucial to think about future paths toward even better testing and rating methods. In particular, there could be tremendous value in defining test approaches that could determine performance maps for heat pumps and air conditioners that could be integrated into simulation tools for estimating more accurate and climate-specific seasonal performance ratings. Future work to develop procedures for testing and performance mapping that minimize test requirements while enhancing reproducibility should be a priority.

REFERENCES

AHRI. (2017). *AHRI Standard 210/240. Performance Rating of Unitary Air-Conditioning and Air-Source Heat Pump Equipment*. Air-Conditioning, Heating, and Refrigeration Institute.

AHRI. (2020). *AHRI Standard 210/240-2023. Performance Rating of Unitary Air-Conditioning & Air-Source Heat Pump Equipment*. Air-Conditioning, Heating, and Refrigeration Institute.

ASHRAE. (2010). *ANSI/ASHRAE Standard 116-2010. Methods of Testing for Rating Seasonal Efficiency of Unitary Air Conditioners and Heat Pumps*. American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE. (2019). *ANSI/ASHRAE Standard 37-2009 (RA 2019). Methods of Testing for Rating Electrically Driven Unitary Air-Conditioning and Heat Pump Equipment*. American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Cheng, L., Dhillon, P., Horton, W. T., & Braun, J. E. (2021). Automated laboratory load-based testing and performance rating of residential cooling equipment. *International Journal of Refrigeration*, *123*, 124–137. https://doi.org/10.1016/j.ijrefrig.2020.11.016

Cheng, L., Patil, A., Dhillon, P., Braun, J. E., & Horton, W. T. (2018). Impact of Virtual Building Model and Thermostat Installation on Performance and Dynamics of Variable-Speed Equipment during Load-based Tests. *International Refrigeration and Air Conditioning Conference*, Paper 2078. https://docs.lib.purdue.edu/iracc/2078

Cremaschi, L., & Perez Paez, P. (2017). Experimental feasibility study of a new load-based method of testing for light commercial unitary heating, ventilation, and air conditioning (ASHRAE RP-1608). *Science and Technology for the Built Environment*, *23*(7), 1178–1188. https://doi.org/10.1080/23744731.2016.1274628

CSA. (2019). *CSA EXP07:19 Load-based and climate-specific testing and rating procedures for heat pumps and air conditioners*.

Dhillon, P., Cheng, L., Horton, W. T., & Braun, J. E. (2022a). Laboratory Load-Based Testing and Performance Rating of Residential Heat Pumps in Heating Mode. *Science and Technology for the Built Environment*.

Dhillon, P., Horton, W. T., & Braun, J. E. (2022b). *AHRI 8026 - Repeatability and Reproducibility Assessment of CSA EXP07:19 and AHRI 210-240:2023*. Air Conditioning, Heating, and Refrigeration Institute.

Dhillon, P., Horton, W. T., & Braun, J. E. (2022c). Comparison of Residential Heat Pump Heating Seasonal Performance based on Load-Based and Steady-State Testing Methodologies. *ASHRAE Winter Conference*.

Dhillon, P., Horton, W. T., & Braun, J. E. (2021a). Comparison of Steady-State and Dynamic Load-Based Performance Evaluation Methodologies for a Residential Air Conditioner. *International Refrigeration and Air Conditioning Conference*, Paper 2172. https://docs.lib.purdue.edu/iracc/2172

Dhillon, P., Horton, W. T., & Braun, J. E. (2021b). Demonstration of a Load-Based Testing Methodology for Rooftop Units with Integrated Economizers. *International Refrigeration and Air Conditioning Conference*, Paper 2174. https://docs.lib.purdue.edu/iracc/2174

Dhillon, P., Horton, W. T., & Braun, J. E. (2021c). Load-Based Testing Methodology for Evaluating Advanced Heat Pump Control Design. *13th IEA Heat Pump Conference*, 1863–1874.

Dhillon, P., Kim, D., Horton, W. T., & Braun, J. E. (2022d). Repeatability Assessment of Load-Based Testing Methodology for Residential Air Conditioning Equipment. *ASHRAE Winter Conference*.

Dhillon, P., Patil, A., Cheng, L., Braun, J. E., & Horton, W. T. (2018). Performance Evaluation of Heat Pump Systems Based on a Load-based Testing Methodology. *International Refrigeration and Air Conditioning Conference*, Paper 2077. https://docs.lib.purdue.edu/iracc/2077

Dhillon, P., Welch, D., Butler, B., Horton, W. T., & Braun, J. E. (2021d). Validation of a Load-Based Testing Method for Characterizing Residential Air-Conditioner Performance. *International Refrigeration and Air Conditioning Conference*, Paper 2257. https://docs.lib.purdue.edu/iracc/2257

Dhillon, P., Welch, D., Butler, B., Horton, W. T., & Braun, J. E. (2022e). Validation of a Load-Based Testing Methodology for Residential Heat Pump Performance Characterization in Heating Mode. *International Refrigeration and Air Conditioning Conference*, Paper 2524.

Hjortland, A. L., & Braun, J. E. (2019). Load-based testing methodology for fixed-speed and variable-speed unitary air conditioning equipment. *Science and Technology for the Built Environment*, *25*(2), 233–244. https://doi.org/10.1080/23744731.2018.1520564

Kim, D., Dhillon, P., Horton, W. T., & Braun, J. E. (2022). Thermostat Environment Emulator Design Update and Assessment for Load Based Testing Methodology. *International Refrigeration and Air Conditioning Conference*, Paper 2522.

Ma, J., Dhillon, P., Horton, W. T., & Braun, J. E. (2021). Heat-Pump Control Design Performance Evaluation using Load-Based Testing. *International Refrigeration and Air Conditioning Conference*, Paper 2173. https://docs.lib.purdue.edu/iracc/2173

Patil, A., Hjortland, A. L., Cheng, L., Dhillon, P., Braun, J. E., & Horton, W. T. (2018). Load-Based Testing to Characterize the Performance of Variable-Speed Equipment. *International Refrigeration and Air Conditioning Conference*, Paper 2076. https://docs.lib.purdue.edu/iracc/2076

ACKNOWLEDGEMENT

This work was supported by AHRI, NRCan, NEEA, and PG&E. The authors would like to acknowledge Mark Baines, Kevin McFadden, and the testing staff at UL and PG&E labs for their help and support in the experimental work. The authors are also thankful for the guidance and feedback from AHRI Unitary Alternate Rating Working Group (UAR WG) members.