On the Understandability of Temporal Properties Formalized in Linear Temporal Logic, Property Specification Patterns and Event Processing Language

Christoph Czepa and Uwe Zdun

APPENDIX

Index Terms—Controlled Experiment, Understandability, Temporal Property, Linear Temporal Logic, Property Specification Patterns, Complex Event Processing, Event Processing Language

APPENDIX A
ANALYSIS

A.1 Data Set Preparation

The first experiment run considered the overall response time per participant only. In the second run, we introduced a more fine-grained approach for time tracking that works on a per task basis. Unfortunately, a small number of the participants of the second experiment run failed to perform the time tracking per task correctly. Moreover, one participant used an answer sheet of a different group, and a few students already participated in the first experiment run in the course of their previous studies. Due to the large number of remaining observations, we decided to drop the incomplete and potentially unreliable data of those participants. All dropped participants are summarized in Table 1.

TABLE 1
Summary of dropped participants

Group	Correctness	Response Time	Course	Reason
PSP	43.9 %	-	DSE	Time records missing completely
PSP	16.3 %	22.4 minutes	DSE	Suspicious time record for one task (10 second duration)
PSP	43.0 %	42.0 minutes	ASE	Already participated in first experiment run
PSP	77.8 %	40.1 minutes	ASE	Already participated in first experiment run
LTL	18.1 %	-	DSE	Time records missing for three tasks
LTL	14.8 %	-	ASE	Time records missing for one task; Already participated in first experiment run
LTL	31.7 %	-	DSE	Time records missing for one task
LTL	22.2 %	47.7 minutes	ASE	Already participated in first experiment run
LTL	-	50.0 minutes	DSE	Wrong answer sheet used
EPL	50.6 %	-	DSE	Time records missing for three tasks

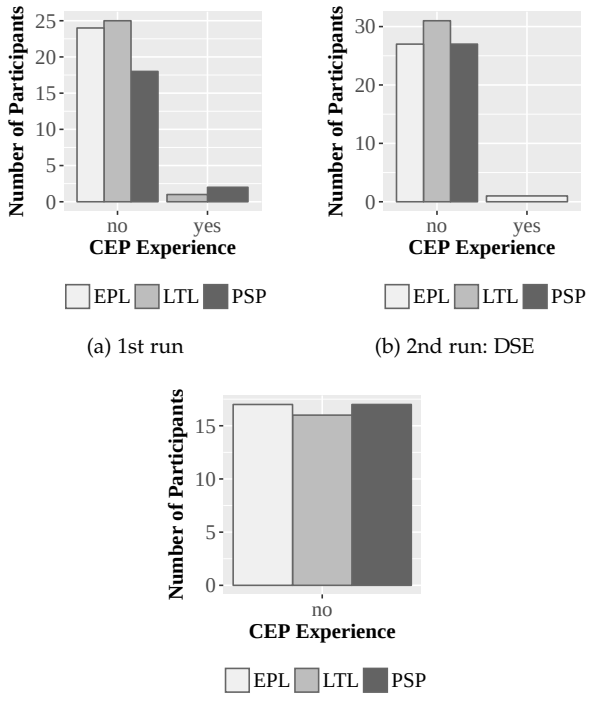
A.2 Descriptive Statistics

The purpose of this section is to present the collected data (cf. Czepa & Zdun [1]) with the help of descriptive statistics. First, we analyze the previous knowledge and experience of the participants. By comparing the previous knowledge and other features (e.g., age of the participants) of the different groups, we try to find out whether the random allocation of participants to groups has led to balanced groups or not. Following this, we will use descriptive statistics to analyze the dependent variables.

A.2.1 Descriptive Statistics of Previous Knowledge, Experience and Other Features of Participants

Figure 1 shows a bar chart of the participants' previous knowledge of Complex Event Processing (CEP). The distribution between the groups is relatively well-balanced. Overall, only a very few participants are experienced with CEP. Figure 2 shows a bar chart of the participants' previous knowledge of logical formalisms (e.g., first-order logic) in general. Again, the distribution between the groups is relatively well-balanced. Interestingly, the students in ASE seem to be less experienced with logical formalisms than the DSE students in the second experiment run. A possible reason for this might be that more time has passed between attending the respective lectures introducing the formalisms for master students in ASE than for bachelor students in DSE and the diverse background of our master students (i.e., coming from various faculties and countries with different curricula).

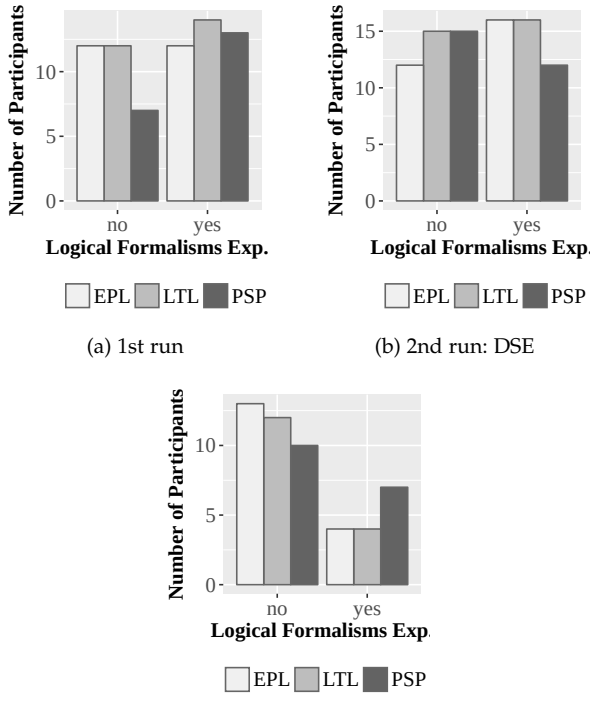
• *The authors are with the Research Group Software Architecture, Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria*
E-mail: christoph.czepa@univie.ac.at, uwe.zdun@univie.ac.at



(a) 1st run (b) 2nd run: DSE

(c) 2nd run: ASE

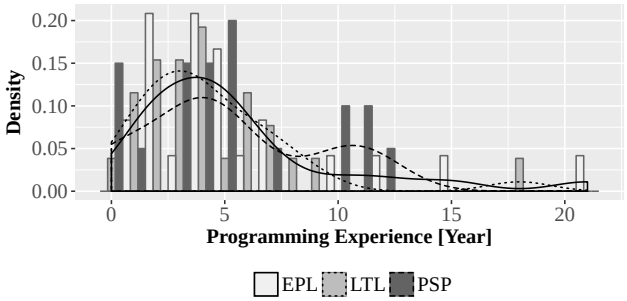
Fig. 1. Bar charts of the participants' experience with Complex Event Processing per group and experiment run



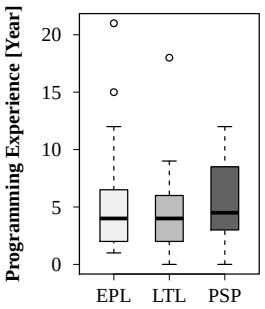
(a) 1st run (b) 2nd run: DSE

(c) 2nd run: ASE

Fig. 2. Bar charts of the participants' experience with logical formalisms per group and experiment run



(a) Kernel density plot

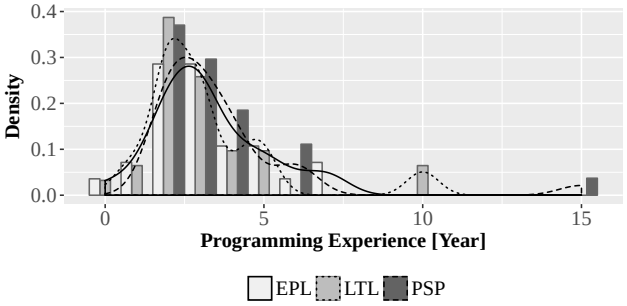


(b) Box plot

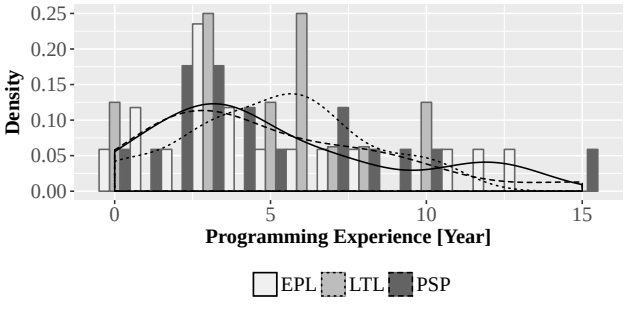
Fig. 3. Kernel density plot and box plot of the participants' programming experience per group in the first experiment run

Next, we investigate the participants' programming experience and work experience in the software industry. Figure 3 shows a kernel density plot and box plot of the programming experience per group in the first experiment run. The peak density of all groups is at about 3 to 4 years of programming experience. Another peak is at about 11 years in the PSP group. Both the EPL and LTL group have a small amount of participants that have 15 and more years of programming experience (shown as outliers in the box plot) while the PSP group has a slightly larger number of participants that have between 10 and 13 years of experience in programming. According to the plots, the participants of the PSP group seem to be slightly more experienced in programming. Figure 4 contains a kernel density plot and box plot of the participants' programming experience in the second experiment run. DSE participants have the peak density in all groups at about 3 years. Only a very few participants have more than 7 years of programming experience in all groups. According to these plots, the distribution is similar in all three experiment groups. In ASE, we can observe a difference in the central tendency in the LTL group which has its peak density at about 6 years, whereas the peak density of the two other groups is at about 4 years. Above 10 years of experience occurs only in the EPL and PSP groups. Thus, those groups contain a few highly experienced programmers, and the LTL group appears to be slightly more experienced on the average.

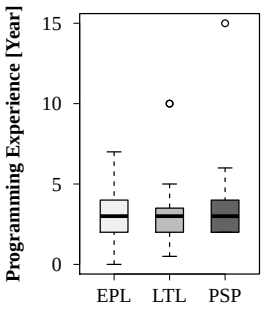
Figure 5 shows the participants' experience with regard to working in the software industry in the first experiment run. The majority of participants do not have any such work experience at all. Overall, the shapes and peaks of the distributions are rather similar. Some EPL participants have



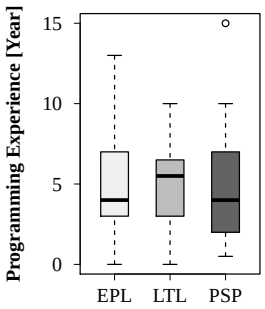
(a) Kernel density plot: DSE



(b) Kernel density plot: ASE



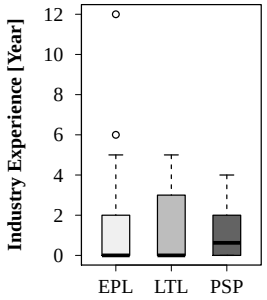
(c) Box plot: DSE



(d) Box plot: ASE

Fig. 4. Kernel density plots and box plots of the participants' programming experience per group in the second experiment run

(a) Kernel density plot



(b) Box plot

Fig. 5. Kernel density plot and box plot of the participants' software industry experience per group in the first experiment run

a higher number of years of experience in comparison to the participants in the other groups (shown as outliers in the box plot). The PSP group has slightly more work experience on the average. In Figure 6, the participants' industrial experience in the second experiment run is shown. The peak density of all groups in DSE is at zero years. In ASE, the LTL group has slightly less working experience than the two other groups apparently.

In the second experiment run, we additionally gathered information regarding the age and gender of the participants. In Figure 7, the participants' age per group is shown. On the average, DSE students of the LTL group are slightly older than their colleagues in the PSP group, and DSE students of the EPL group are younger than their colleagues in the two other groups. Overall, the majority of the DSE participants shares the same age group (20–25). In ASE, the participants of the EPL group are on average slightly older. Moreover, the kernel density plot suggests that there are two age groups, namely younger students (aged 22–27) and older students (aged 30–35). The fraction of female participants is slightly lower in the PSP group in DSE (cf. Figure 8). Overall, the distribution of male and female participants is balanced.

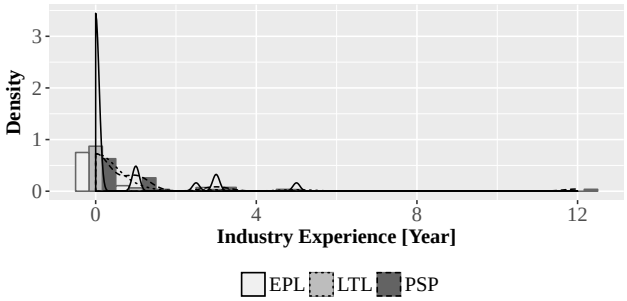
According to the descriptive statistics which indicates merely minor differences, the groups in both experiment runs are similar with regards to previous knowledge, experience, and also with regards to age and gender in the second run. No major differences between the groups are noticeable.

A.2.2 Descriptive Statistics of Dependent Variables

Table 2 contains the number of observations, central tendency measures and dispersion measures of the dependent variables (correctness and response time) per temporal property representation and experiment run. The second experiment run consists of measurements in two courses, namely DSE and ASE. That is, we tested our hypotheses three times, namely in the first experiment run in ASE, and in the second experiment run in DSE and ASE. In all three cases, the PSP group reached the highest mean and median correctness (about 70–75%), followed by the EPL group (about 50–55% correctness) and the LTL group (about 30–35% correctness). The maximum measured response time in the first run is the 90 minutes limit in all groups. In response to this, we reduced the number of tasks in the second run by one (from 10 to 9). In the second run, the maximum response time is 88 minutes. Interestingly, students in the second run in ASE managed to finish on the average about 20–40% faster than their colleagues in the first run which cannot be caused by the removal of a single task alone as the expected response time reduction would be only about 10%. We suspect that this difference is caused by the change from total experiment time recordings in the first experiment

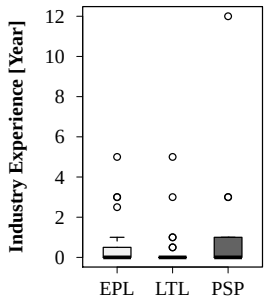
TABLE 2
Number of observations, central tendency and dispersion per group and experiment run

		LTL	PSP	EPL
1st run	Number of observations	26	20	24
	Mean correctness [%]	33.04	69.55	50.70
	Standard deviation [%]	15.39	25.46	28.52
	Median correctness [%]	31.3	78	48.7
	Median absolute deviation [%]	12.79	23.87	42.48
	Min. correctness [%]	5	12.7	10.5
	Max. correctness [%]	63	100	94.7
	Skew (correctness)	0.02	−0.56	0.01
	Kurtosis (correctness)	−0.83	−1.01	−1.61
	Mean response time [min]	69.85	58.25	72.12
	Standard deviation [min]	15.25	20.86	21.47
	Median response time [min]	73	57.50	78.5
	Median absolute deviation [min]	17.05	25.95	17.05
	Min. response time [min]	35	28	11
	Max. response time [min]	90	90	90
	Skew (response time)	−0.44	0.13	−1.44
	Kurtosis (response time)	−0.74	−1.53	1.25
2nd run: DSE	Number of observations	31	27	28
	Mean correctness [%]	32.45	70.55	53.83
	Standard deviation [%]	17.23	20.89	23.04
	Median correctness [%]	31.7	73.70	54.10
	Median absolute deviation [%]	18.09	18.09	23.5
	Min. correctness [%]	6.5	16.30	5.6
	Max. correctness [%]	70.6	97.20	86.70
	Skew (correctness)	0.36	−0.87	−0.37
	Kurtosis (correctness)	−0.62	−0.11	−0.86
	Mean response time [min]	51.03	36.65	43.80
	Standard deviation [min]	14.95	14.18	14.71
	Median response time [min]	51	33.05	42.76
	Median absolute deviation [min]	13.42	15.25	13.2
	Min. response time [min]	19	17.35	23
	Max. response time [min]	88	63.08	84.63
	Skew (response time)	0.25	0.56	0.81
	Kurtosis (response time)	−0.10	−1.10	0.38
2nd run: ASE	Number of observations	16	17	17
	Mean correctness [%]	36.42	72.41	54.4
	Standard deviation [%]	17.32	18.17	21.06
	Median correctness [%]	38.60	71.9	53.70
	Median absolute deviation [%]	9.71	18.09	17.35
	Min. correctness [%]	3.7	33.50	8.9
	Max. correctness [%]	67.6	100	87.6
	Skew (correctness)	−0.08	−0.23	−0.32
	Kurtosis (correctness)	−0.63	−0.78	−0.70
	Mean response time [min]	55.32	39.12	44
	Standard deviation [min]	11.51	8.95	15.33
	Median response time [min]	53.15	39.5	44.83
	Median absolute deviation [min]	11.48	9.64	19.74
	Min. response time [min]	35.5	23.47	23
	Max. response time [min]	78	52.93	70.5
	Skew (response time)	0.27	−0.23	0.30
	Kurtosis (response time)	−0.90	−1.19	−1.35

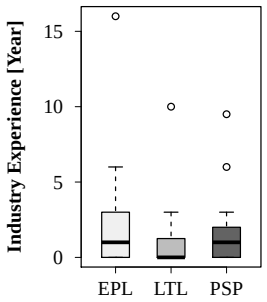


(a) Kernel density plot: DSE

(b) Kernel density plot: ASE



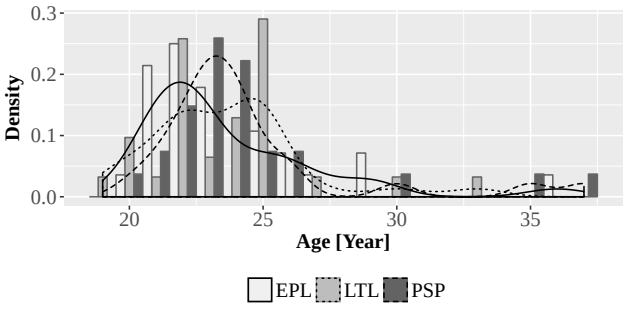
(c) Box plot: DSE



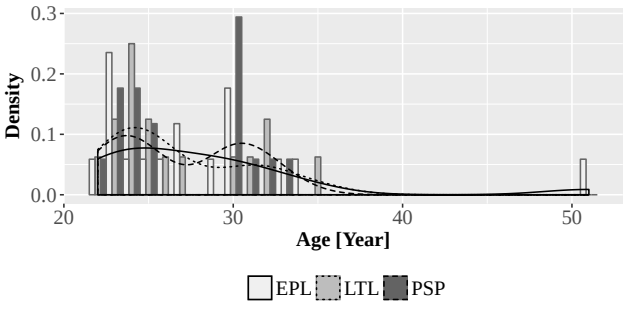
(d) Box plot: ASE

Fig. 6. Kernel density plots and box plots of the participants' software industry experience per group in the second experiment run

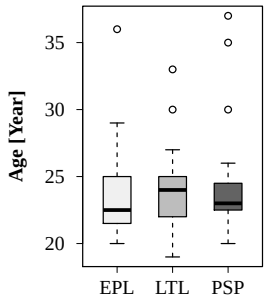
run to per task time recordings in the second experiment run, and the late assignment of participants to groups at the beginning of the experiment session in the first run. Obviously, the time recordings of the participants in the first experiment run included times such as pauses, task switching times, and times spent on consulting the accompanying documents that are not directly related to solving a specific task. In the first experiment run the participants had to be prepared for all three representations, and the experiment group was assigned at the beginning at the experiment session. Up to this point in time, the participants did not know to which experiment group they were assigned to. That is, once it became clear which of the three approaches must be applied, the participants revisited the learning material related to the assigned representation intensely. In the second experiment run, group assignment was clear beforehand, so this initial consulting of the info material did not take place in a comparable intensity. Furthermore, the mean (72.12 minutes) and median response times (78.5 minutes) of the EPL group are longer than those of the LTL



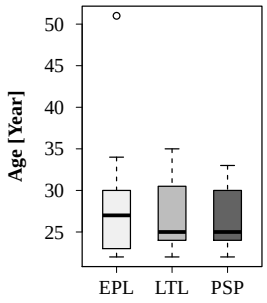
(a) Kernel density plot: DSE



(b) Kernel density plot: ASE

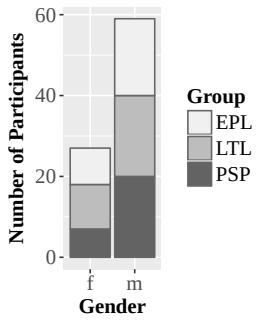


(c) Box plot: DSE

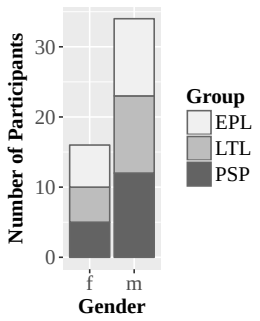


(d) Box plot: ASE

Fig. 7. Kernel density plots and box plots of the participants' age per group in the second experiment run



(a) DSE



(b) ASE

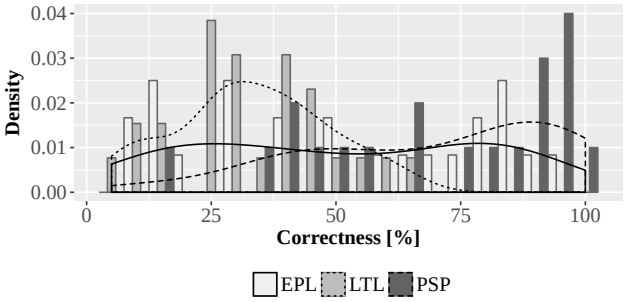
Fig. 8. Bar charts of the participants' gender per group in the second experiment run

group (69.85 minutes mean and 73 minutes median) in the first run. With regard to the hypotheses of this experiment, the response time measurements in the first experiment run are an unexpected result since we expected that the response times in the EPL group would be faster than in the LTL group. In contrast, the EPL group has a faster response time than the LTL group in the second run. We suspect that this effect could have been caused by the task design which contained truth value states in the answer choices that are not part of the EPL temporal property definition. Originally (i.e., at the time the first run was completed, and before the second run was carried out), we thought that there might have been a bias present in the first experiment run in favor of the EPL group, because wrong answer choices could have been potentially easier to identify by the EPL participants. However, these answer choices seemingly rather confused the participants than helped them. During the the first experiment run, EPL participants repeatedly asked whether there is an error in the exercise or whether it can be really that easy to solve it. Due to their confusion, EPL participants spent considerable more time on solving the tasks in the first experiment run. The skew values of the correctness variable are balanced (i.e., close to zero) for the LTL and EPL groups in the first run. That is, the distribution is rather symmetric. The negative PSP correctness skew value (-0.56) suggests that the distribution is left-tailed. A positive value such as the skew of the response time variable in the second run in DSE indicates a right-tailed distribution. Kurtosis, another measure for the shape of a distribution, focuses on the general tailedness of a distribution. A negative kurtosis indicates fat tails, and vice versa, a positive kurtosis indicates skinny tails with a distribution toward the mean. In general, the differences in skew and kurtosis between the groups indicate differences in the shape of their distributions. Consequently, the skew and kurtosis values in Table 2 suggest that there exist changes in distribution of the dependent variables between the groups.

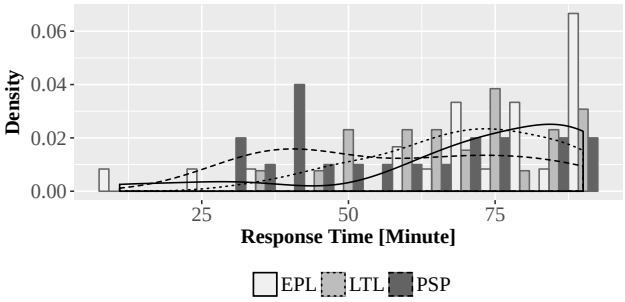
Additionally to the descriptive statistics in Table 2, we will now perform a graphical analysis that is based on kernel density plots and box plots to further study the dependent variables. Kernel density plots are well-suited to visualize the distribution of the data whereas box plots are used to visualize the quartiles and outliers.

In the first experiment run (as shown in Figure 9), the EPL correctness distribution is extremely long-tailed and flat. The distribution of the PSP correctness has its peak close to the maximum and a long left tail. LTL has the steepest correctness distribution with its peak at about 30% and a right tail that already ends at about 75% correctness. The EPL response time distribution has its peak close to the maximum of 90 minutes and a slope until about 50 minutes where the density is already low, but remaining nearly constant from that point on. LTL response time has its peak density at about 75 minutes and a long left tail that ends at about 25 minutes. There are two response time outliers in the EPL group which do not result from measuring errors.

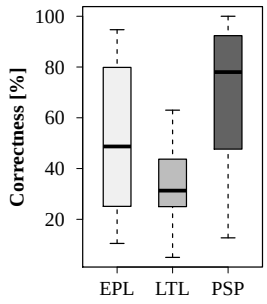
In the second experiment run with DSE participants (as shown in Figure 10), the EPL correctness distribution is still rather flat, but less extreme than in the first run. PSP has its peak correctness at about 90% and a long left tail. The kernel density plot shows the peak correctness in the LTL group



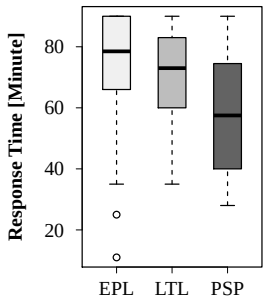
(a) Kernel density plot: Correctness



(b) Kernel density plot: Response time

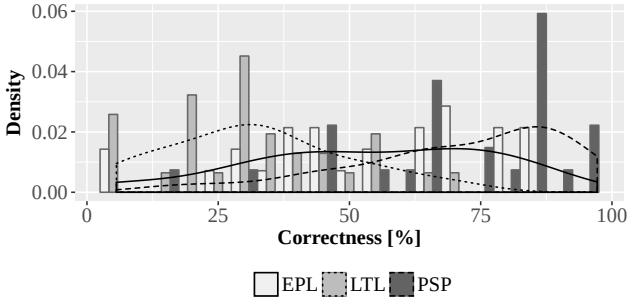


(c) Box plot: Correctness

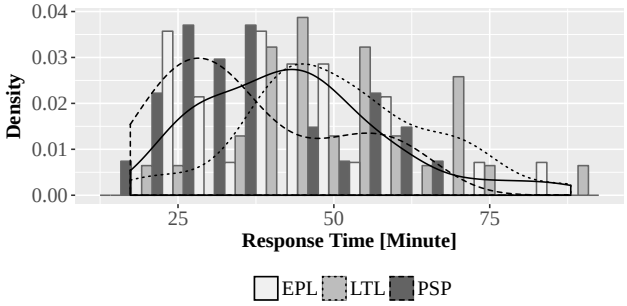


(d) Box plot: Response time

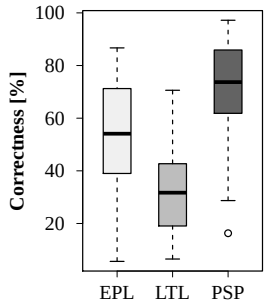
Fig. 9. Kernel density plots and box plots of the participants' overall correctness of the given answers and the overall response time per group in the first experiment run



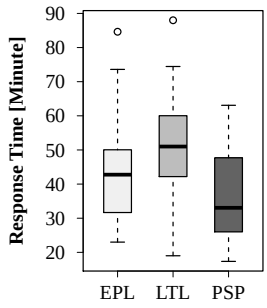
(a) Kernel density plot: Correctness



(b) Kernel density plot: Response time



(c) Box plot: Correctness



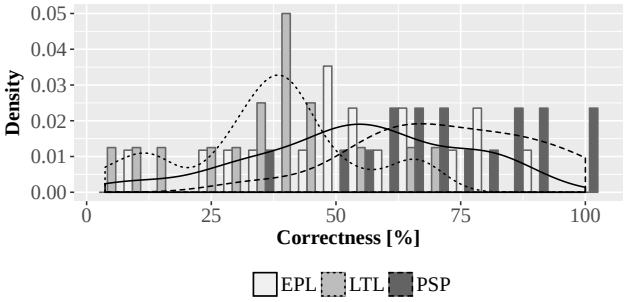
(d) Box plot: Response time

Fig. 10. Kernel density plots and box plots of the DSE participants' overall correctness of the given answers and the overall response time per group in the second experiment run

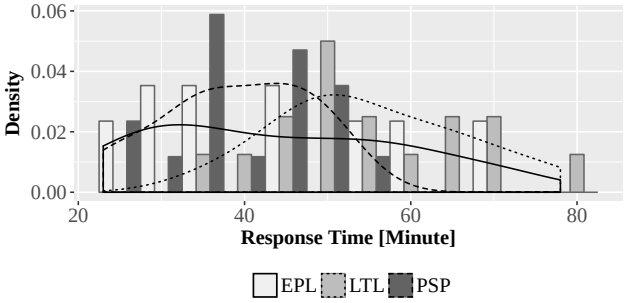
at about 30%, and the right tail of the distribution ends at about 75%. In contrast to the first run, we observe faster response times overall and especially in the EPL group. The peaks of the LTL and EPL response time distributions share nearly the same location at about 45 minutes. Apart from that, the distributions are fairly different because the EPL group has a higher density on the left tail whereas the LTL group has a higher density on the right tail. In the PSP group, the highest density is located at about 25–30 minutes with another smaller peak at about 55–60 minutes. There is a single correctness outlier in the PSP group, and there are two response time outliers, one in the LTL group and another in the EPL group. Since those outliers are not caused by measuring errors, we see no reason for exclusion.

In the second experiment run with ASE participants (as shown in Figure 11), the peak density of the LTL group correctness is located at about 35–40%. Two tiny peaks can be found at about 10–15% and 60–65% correctness. The peak density of the PSP group is located at 60–65%, and the density drops merely slowly on the right tail which indicates a high level of correctness in this group. The EPL group has its peak correctness at about 55%, and the shape indicates that the right tail has a slightly higher density than the left tail. Like in Figure 9 and Figure 10, the response time distribution is relatively flat in the EPL group. The highest density can be found at about 30–35 minutes. From that point on, the density is slowly decreasing. PSP has the steepest response time distribution with its peak at about 45 minutes and higher density on the left tail. In contrast, the LTL response time distribution has a higher density on the right tail, and the peak response time is located at about 50 minutes. There are two correctness outliers in the LTL group, and a single correctness outlier in the EPL group. Again, those are valid measurements, and we see no reason for excluding them.

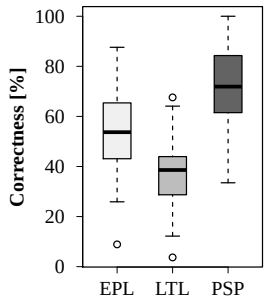
Generally, the distributions look fairly different, which implies unequal variances in the different groups, and there are obvious differences in central tendency. All present



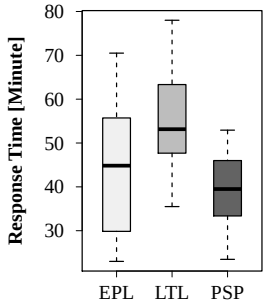
(a) Kernel density plot: Correctness



(b) Kernel density plot: Response time

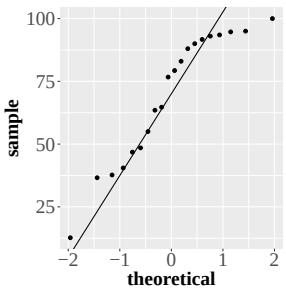


(c) Box plot: Correctness

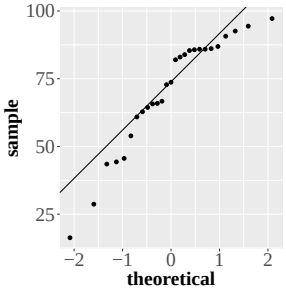


(d) Box plot: Response time

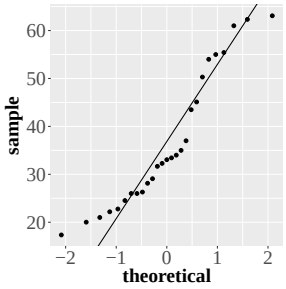
Fig. 11. Kernel density plots and box plots of the ASE participants' overall correctness of the given answers and the overall response time per group in the second experiment run



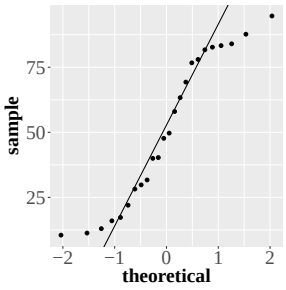
(a) Correctness data of PSP group in the 1st experiment run



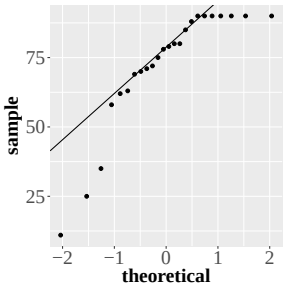
(b) Correctness data of PSP group in the 2nd experiment run (DSE)



(c) Response time data of PSP group in the 2nd experiment run (DSE)



(d) Correctness data of EPL group in the 1st experiment run



(e) Response time data of EPL group in the 1st experiment run

Fig. 12. Normal QQ plots

outliers appear to be valid measurements, so there is not enough evidence to drop them.

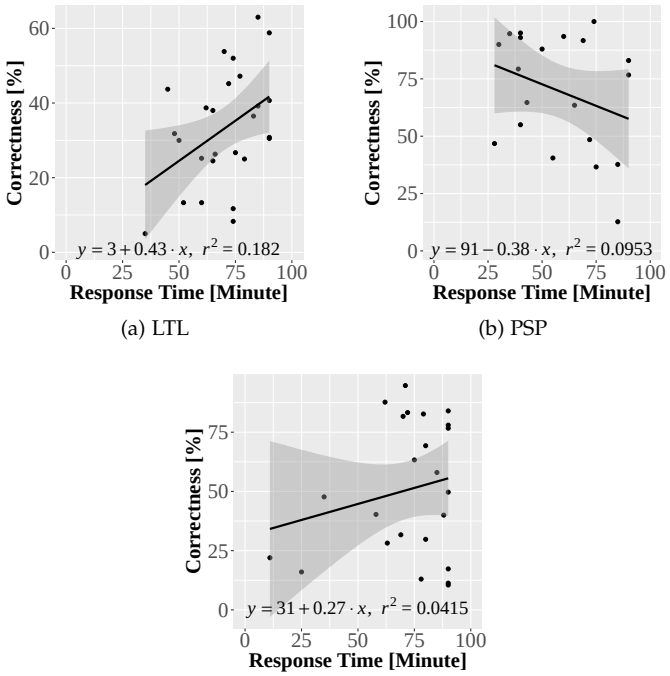
A graphical analysis by normal Q-Q plots and Shapiro-Wilk tests of normality (cf. Table 3) suggest that the univariate normality assumption does not hold in multiple cases. In the following, we discuss the most severe cases. Specifically, the univariate normality assumption does not hold for

- the correctness variable of the PSP group in the first and second (DSE) experiment run (cf. Figure 12 (a) & (b)),
- the response time variable of the PSP group in the second (DSE) experiment run (cf. Figure 12 (c)),
- the correctness and response time variable of the EPL group in the first experiment run (cf. Figure 12 (d) & (e)).

Scatter plots (as shown in Figure 13) and Kendall's rank correlation tau tests (summarized in Table 4) do not indicate any significant correlation of the two dependent variables (i.e., correctness and response time). Please note that the

TABLE 3
Shapiro-Wilk test of normality (* for $\alpha = 0.05$, ** for $\alpha = 0.01$, * for $\alpha = 0.001$)

Group	Dependent Variable	1st Run	2nd Run: DSE	2nd Run: ASE
LTL	Correctness	$W = 0.9782$ $p = 0.8328$	$W = 0.96$ $p = 0.3096$	$W = 0.9598$ $p = 0.6581$
	Response Time	$W = 0.9501$ $p = 0.2326$	$W = 0.976$ $p = 0.696$	$W = 0.9838$ $p = 0.9867$
PSP	Correctness	$W = 0.902$ $p = 0.045$ *	$W = 0.9062$ $p = 0.0186$ *	$W = 0.9725$ $p = 0.8606$
	Response Time	$W = 0.9216$ $p = 0.1063$	$W = 0.9047$ $p = 0.0172$ *	$W = 0.9598$ $p = 0.6277$
EPL	Correctness	$W = 0.9109$ $p = 0.0369$ *	$W = 0.9539$ $p = 0.2473$	$W = 0.9753$ $p = 0.9023$
	Response Time	$W = 0.7947$ $p = 0.0002$ ***	$W = 0.9402$ $p = 0.112$	$W = 0.9314$ $p = 0.2298$



(a) LTL (b) PSP

(c) EPL

Fig. 13. Scatter plots of response time vs. correctness in first experiment run with linear trend lines, 95% confidence regions, and coefficients of determination (r^2)

TABLE 4
Kendall's rank correlation tau (* for $\alpha = 0.05$, ** for $\alpha = 0.01$, * for $\alpha = 0.001$)

Group	1st Run	2nd Run: DSE	2nd Run: ASE
LTL	$\tau = 0.2638$ $z = 1.8589$ $p = 0.0631$	$\tau = -0.0693$ $z = -0.5445$ $p = 0.5861$	$\tau = 0.1833$ $T = 71$ $p = 0.3502$
PSP	$\tau = -1.2691$ $z = -1.2691$ $p = 0.2044$	$\tau = -0.0029$ $z = -0.0209$ $p = 0.9834$	$\tau = -0.155$ $z = -0.8658$ $p = 0.3866$
EPL	$\tau = -0.0151$ $z = -0.1006$ $p = 0.9198$	$\tau = 0.15$ $z = 1.1263$ $p = 0.26$	$\tau = 0.0441$ $T = 71$ $p = 0.8393$

scatter plots in the second experiment run reveal a similar picture, so they are omitted intentionally in Figure 13.

APPENDIX B
STATISTICAL INFERENCE

The multivariate analysis of variance (MANOVA) is a suitable statistical inference procedure in the presence of two dependent variables. However, necessary assumptions must be met. Please note that we will not discuss each and every assumption or report its violation if a specific other (more elementary) assumption already indicates a violation that hinders any meaningful application of the method on the given data set. Both the graphical analysis (by kernel density plots and normal Q-Q plots) and Shapiro-Wilk tests of the data indicate that the univariate normality assumption does not hold in multiple cases. The linearity assumption demands that all of the dependent variables are linearly related to each other, but scatter plots and Residuals vs. Fitted plots suggest that the linearity assumption is not met by the data sufficiently. As a result, the power of the multivariate and parametric MANOVA test might be affected, and its results would be unreliable. Multivariate and parametric testing might lead to unreliable results due to unsatisfied model assumptions, so we fall back to univariate non-parametric testing. The univariate non-parametric Kruskal-Wallis test is strongly affected by unequal variances (cf. Kitchenham et al. [2]), so its result might be not reliable because the kernel density plots of the data show distributions that look different in many cases which implies unequal variances in the different groups.

TABLE 5
Cliff's d (first experiment run), one-tailed with confidence intervals calculated for $\alpha = 0.05$ (cf. Cliff [3] and Rogmann [4]), adjusted p-values (cf. Benjamini & Hochberg [5]) [Level of significance: * for $\alpha = 0.05$, ** for $\alpha = 0.01$, *** for $\alpha = 0.001$], and effect size magnitudes (cf. Kitchenham et al. [2])

		PSP/LTL	PSP/EPL	EPL/LTL
Correctness	$p_1 = P(X > Y)$	0.8769	0.7021	0.6715
	$p_2 = P(X = Y)$	0	0.0042	0
	$p_3 = P(X < Y)$	0.1231	0.2938	0.3285
	d	-0.7539	-0.4083	-0.343
	s_d	0.1097	0.1575	0.162
	z	-6.8699	-2.5933	-2.1157
	CI low	-0.8847	-0.633	-0.5789
	CI high	-0.513	-0.1203	-0.0539
	p	8.8×10^{-9}	0.0065	0.0198
	FDR adjusted p	5.3×10^{-8}	0.0195	0.0297
	level of significance	***	*	*
	effect size magnitude	large	medium	medium
Response Time	$p_1 = P(X > Y)$	0.3115	0.2833	0.564
	$p_2 = P(X = Y)$	0.0442	0.0417	0.0577
	$p_3 = P(X < Y)$	0.6442	0.675	0.3782
	d	0.3327	0.3917	-0.1859
	s_d	0.1693	0.1641	0.164
	z	1.9649	2.3874	-1.1336
	CI low	0.0312	0.0931	-0.4376
	CI high	0.5787	0.6256	0.0928
	p	0.0279	0.0108	0.1313
	FDR adjusted p	0.0335	0.0216	0.1313
	level of significance	*	*	-
	effect size magnitude	medium	medium	-

As a consequence, we use *Cliff's delta* (cf. Cliff [3] and Rogmann [4]), a robust non-parametric test that is unaffected by change in distribution, non-normal-data and possible non-stable variance. The results of the test are shown in Table 5 for the first experiment run and Table 6 for the second experiment run where

- p_1 represents the probability that a subject chosen from group X has a higher value than a randomly chosen subject from group Y,
- p_2 reflects the probability that a subject chosen from group X has an equal value to a randomly chosen subject from group Y,
- p_3 is the probability of superiority of Y over X,
- d denotes Cliff's delta for independent groups (i.e., the difference between the probability that a randomly chosen Y measurement has a higher value

TABLE 6
Cliff's d (second experiment run), one-tailed with confidence intervals calculated for $\alpha = 0.05$ (cf. Cliff [3] and Rogmann [4]), adjusted p-values (cf. Benjamini & Hochberg [5]) [Level of significance: * for $\alpha = 0.05$, ** for $\alpha = 0.01$, *** for $\alpha = 0.001$], and effect size magnitudes (cf. Kitchenham et al. [2])

		PSP/LTL	PSP/EPL	EPL/LTL
Correctness in DSE	$p_1 = P(X > Y)$	0.902	0.709	0.7661
	$p_2 = P(X = Y)$	0	0.0053	0.0012
	$p_3 = P(X < Y)$	0.098	0.2857	0.2327
	d	-0.8041	-0.4233	-0.5334
	s_d	0.0847	0.1388	0.1275
	z	-9.4883	-3.0494	-4.1823
	CI low	-0.9053	-0.6238	-0.7107
	CI high	-0.6163	-0.1706	-0.2924
	p	1.5×10^{-13}	0.0018	5.0×10^{-5}
	FDR adjusted p	8.9×10^{-13}	0.0027	0.0002
	level of significance	***	**	***
	effect size magnitude	large	medium	large
Response Time in DSE	$p_1 = P(X > Y)$	0.2473	0.3585	0.3502
	$p_2 = P(X = Y)$	0.0024	0.0027	0.0023
	$p_3 = P(X < Y)$	0.7503	0.6389	0.6474
	d	0.503	0.2804	0.2972
	s_d	0.1345	0.153	0.1452
	z	3.7393	1.8324	2.0474
	CI low	0.251	0.0135	0.0432
	CI high	0.6911	0.51	0.5151
	p	0.0002	0.0363	0.0226
	FDR adjusted p	0.0004	0.0363	0.0271
	level of significance	***	*	*
	effect size magnitude	large	medium	medium
Correctness in ASE	$p_1 = P(X > Y)$	0.9154	0.7405	0.7427
	$p_2 = P(X = Y)$	0	0	0.0037
	$p_3 = P(X < Y)$	0.0846	0.2595	0.2537
	d	-0.8309	-0.481	-0.489
	s_d	0.0968	0.1691	0.1761
	z	-8.5879	-2.8448	-2.777
	CI low	-0.9352	-0.7108	-0.7248
	CI high	-0.5937	-0.1585	-0.1506
	p	5.3×10^{-10}	0.0038	0.0046
	FDR adjusted p	3.2×10^{-9}	0.0069	0.0069
	level of significance	***	**	**
	effect size magnitude	large	large	large
Response Time in ASE	$p_1 = P(X > Y)$	0.125	0.4187	0.2794
	$p_2 = P(X = Y)$	0.0037	0	0.0037
	$p_3 = P(X < Y)$	0.8713	0.5813	0.7169
	d	0.7463	0.1626	0.4375
	s_d	0.1172	0.2063	0.1814
	z	6.3672	0.7883	2.4124
	CI low	0.4852	-0.1854	0.0972
	CI high	0.8852	0.4744	0.6862
	p	2.2×10^{-7}	0.2	0.011
	FDR adjusted p	6.5×10^{-7}	0.2182	0.0132
	level of significance	***	-	*
	effect size magnitude	large	-	large

than a randomly chosen X measurement and the probability for the opposite),
- s_d is the unbiased sample estimate of the delta standard deviation,
- z is the z-score of Cliff's delta, and
- $(CI\ low,\ CI\ high)$ denotes the confidence interval.

Multiple testing ($n = 6$ because of the two dependent variables and three treatments) requires us to lower the significance level in order to avoid Type I errors (i.e., detection of an effect that is not present). As a classical and widespread method, the Bonferroni correction suggests to lower the alpha value to $\alpha = \frac{0.05}{6} = 0.008\dot{3}$, but the method is also known to skyrocket Type II errors (i.e., failing to detect an effect that is present). As an alternative that is more robust against Type II errors, we consider FDR (False Discovery Rate) adjusted p-values (cf. Benjamini & Hochberg [5]). According to these FDR adjusted p-values, there is evidence for the rejection of the null hypotheses of this study.

In the first experiment run (cf. Table 5), almost all test results are significant which suggests a rejection of $H_{0,1}$ and $H_{0,2}$. $H_{0,3}$ can only be rejected on basis of the correctness variable since the test result does not indicate any significant difference in the response times of the EPL and LTL group. Moreover, the results suggest that the difference in terms of correctness between the PSP and LTL group are highly significant with a large effect size magnitude. All remaining significant test results of the first experiment run show a medium-sized effect.

In the second experiment run (cf. Table 6), the majority of the test results is significant. Only one test, namely the PSP/EPL response time with ASE participants, has no significant result, which means that $H_{0,2}$ (in ASE) can only be rejected on basis of the correctness result. All other test results are ranging from significant ($\alpha = 0.05$) to highly significant ($\alpha = 0.001$) which suggests a rejection of the null hypotheses. Moreover, all significant results show a large or medium effect size magnitude. It is striking that all PSP/LTL test results are highly significant with a large-sized effect.

The statistics software *R*[1] was used for all statistical analyses. In particular, we used the following libraries in the course of our statistical evaluation: *biotools* [6], *car* [7], *ggplot2* [8], *mvnormtest* [9], *mvoutlier* [10], *orddom* [4], *psych* [11], *usdm* [12].

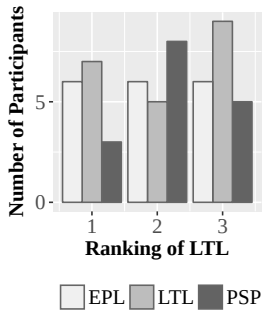
APPENDIX C
ANALYSIS OF QUALITATIVE DATA

In addition to the controlled experiment, we invited the participants of the first experiment run to share their thoughts with regards to the following two tasks:

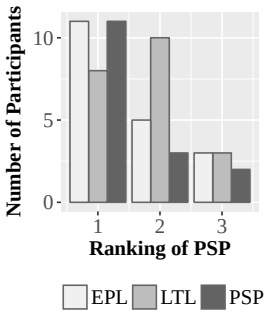
1) *"Please rank the languages according to your preference and state reasons for this ranking."*, and
2) *"Please discuss for which sort of users each language is (not) appropriate and why."*.

The purpose of this survey was to assess the participants' (subjective) preference towards a specific temporal property representation. By that, we tried to gain insights into the users' acceptance of the tested temporal property representations. Please note that we did not replicate this survey in the second experiment run intentionally, because we wanted to avoid the (for this survey necessary) cross-contamination of treatments to improve the validity of the controlled experiment in the second run. Our analysis of the textual answers of the participants has been inspired by the summative content analysis approach [13]. Since the majority of answers given by the participants is very short and in note form, running a full-blown summative content analysis, which usually focuses on journal manuscripts or specific content in textbooks, is impossible. Nevertheless, it is possible to use the core idea of the technique, namely the counting of occurrences of identified keywords and the interpretation of the context associated with the use of the

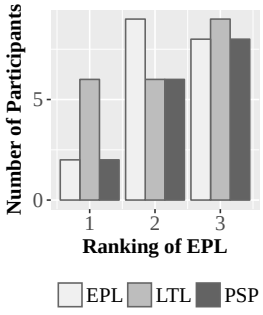
1. https://www.r-project.org/



(a) Personal preference ranking of LTL

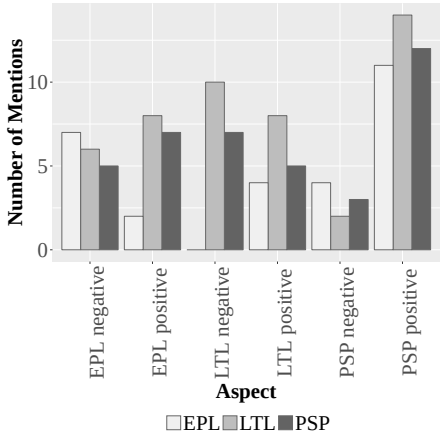


(b) Personal preference ranking of PSP

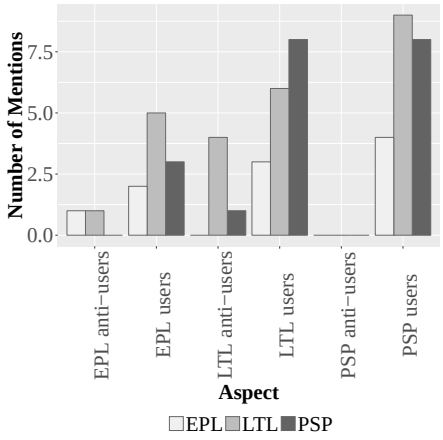


(c) Personal preference ranking of EPL

Fig. 14. Bar charts of the participants' personal preference ranking of the three approaches in the first experiment run



(a) Positive and negative aspects



(b) Potential users and anti-users

Fig. 15. Bar chars of the number of participants mentioning specific aspects of the temporal property representations per group in the first experiment run

word or phrase. In the following, we present the results of this analysis.

Figure 14 shows the personal preference ranking of the tested temporal property representations per group. While the LTL ranking does not show any clear trends, the ranking of the PSP representations indicates a trend towards the first place and the EPL representation towards the third place.

Figure 15 (a) shows a bar chart that contains the number of users that are positive or negative towards a specific temporal property representation. In Figure 15 (b), the number of mentions of user groups for which a specific temporal property representation is considered to be well-suited (i.e., users) or rather problematic (i.e., anti-users) is shown. In all three groups, positive mentions of PSP are dominant. Moreover, the count of mentioned PSP users is overall higher than for the other representations. There has not been a single mention of a user group that should or could not use PSP.

A detailed summary of the mentioned positive and negative aspects, and users and anti-users, is shown in Table 7 and Table 8, respectively. All groups mentioned in the same extent and relatively often (8 times) that PSP is easy to understand. Also the temporal scopes (e.g., After ... until ...) that are present in PSP were mentioned positively. Interestingly, participants other than those of the EPL group relatively often considered EPL as clear and easy to use, while EPL participants apparently did not. Also the separation of concerns in EPL (i.e., through several temporal queries that contain the truth value state change as well) was considered to be a positive aspect of EPL by the LTL group (7 mentions) and the PSP group (2 mentions). Some user comments contradict the comments of other users. For example, one EPL participant stated that EPL is for advanced users while another stated that it is suitable for novice users. Neither participant mentioned any potential anti-user of the PSP representation.

APPENDIX D
THREATS TO VALIDITY

D.1 Threats to Internal Validity

The internal validity is concerned with the causal relationship of independent variables and dependent variables. Threats to internal validity are unknown or unobserved variables that might have an influence on the outcome of the experiment. Diverse threats to internal validity must be addressed:

- *History effects* refer to events that occur in the environment and change the conditions of a study. The short duration of the study limits the possibility of changes in environmental conditions. Actually, we are not aware of any history effects during the study, but we cannot entirely rule out any such effect, prior

TABLE 7
Summary of mentioned positive and negative aspects per group with number of occurrences (if aggregated)

Aspect	LTL Group	PSP Group	EPL Group
LTL positive	can handle all cases, can express everything, powerful logic, most expressive, fleshed-out, many operators, most robust, clear, formal, easy	very/most powerful (2), easy (2), better syntax than PSP, operators for detailed formulas, clear	readability (2), less complicated, easy, clear
LTL negative	hard to read and/or understand (3), nesting (3), complex (2), confusing (2), long formulas are hard to understand	operators are hard to understand, nesting, long formulas are hard to understand, unintuitive, complicated, difficult, complex, hard to comprehend	-
PSP positive	easy (8), scopes (6), clear (2), intuitive (2), self-explaining, very powerful, very logical, high-level	easy (8), scopes (5), intuitive (2), mapping to natural language, common sense, sound set of operators, compact	easy (8), scopes (4), least keywords, most understandable, precise, most readable, close to natural language
PSP negative	complex, operators are hard to understand	insufficient preparation material, scopes hard to understand, confusing, no clear understanding to which state it changes	hard to understand (2), complex keywords, complicated
EPL positive	queries/states (7), clear (6), easy (3)	easy (5), queries/states (2), clear, operators	queries/states (2)
EPL negative	complicated due to multiple queries (2), too many operators, operator precedence, too few operators, event-based character, less powerful	difficult (2), too simple, too complicated tasks, sometimes confusing, complex	complicated (3), demanding (2), poor readability, poor logic, complex

TABLE 8
Summary of mentioned user and anti-user aspects per group with number of occurrences (if aggregated)

Aspect	LTL Group	PSP Group	EPL Group
LTL users	users with basic knowledge, software engineers, expert programmers, developers, programming background, programmers / mathematicians, physicists with prior logic background	experts with many years of experience / experienced users (3), software developers (2), modeling user, users performing model checking, all users including programmers and admins, enduser	mainstream users, experienced users, bank employees after training
LTL anti-users	users with minimal programming experience, all, endusers, economists / simple users without without logical background	project managers	-
PSP users	users with minimal to no experience, diverse users, software architects, in executive presentations, people working with large systems, endusers, small or no programming background, business users, high-level programmer	novice users / beginners (2), programmers (2), modeling user, all kind of users (application users, developers, testing experts), workflow designers, enduser, high-level language for architects	specialized users, unexperienced users, IT-affine people, untrained
PSP anti-users	-	-	-
EPL users	not highly trained staff / novice users (2), software engineers, endusers, database users	interface language between modeling users and programmers, advanced users/modeling users, users experienced with CEP	advanced users, novice users
EPL anti-users	users with minimal programming experience	-	general users

to the study taking place. However, in such a case, it would be extremely unlikely that the scores of one group are more affected than another because of the random allocation of participants to groups.

- *Maturation effects* refer to the impact that time has on an individual. Since the duration of the experiment was very short (max. 90 minutes), maturation effects are considered to be of minor importance.
- *Testing effects* comprise learning effects and experimental fatigue. *Learning effects* were avoided by dropping results of the second run in case of a prior participation in the first run. That is, each person was only tested once. *Experimental fatigue* is concerned with occurrences during the experiment that exhaust the participant either physically or mentally. Neither did we observe any signs of fatigue nor reported any participant any such.
- *Instrumental bias* occurs if the measuring instrument (i.e., a physical measuring device or the actions/assessment of the researcher) changes over time during the experiment. We avoided such effects by using an experimental design that enables an automated and standardized evaluation of the test results.
- *Selection bias* is present if the experimental groups are unequal before the start of the experiment (e.g., severe differences in relevant experience, age, or gender). Usually, selection bias is likely to be more threatening in quasi-experimental research. By using an experimental research design with the fundamental requirement to randomly assignment participants to the different groups of the experiment, we can avoid selection bias to a large extent. In addition, our investigation of the composition of the groups did not indicate any major differences between them.
- *Experimental mortality* is only likely to occur if the experiment lasts for a long time because the chances for dropouts increase (e.g., location change). Consequently, it has not been a problem in our study at all.
- *Diffusion of treatments* occurs if a group of the experiment is contaminated in some way. By design, in the first run of the experiment, making information about all three temporal property languages available to every participant was necessary for the survey. That is, we accepted the risk of cross-contamination intentionally in the first experiment run. In the second experiment run, the survey was not replicated to avoid the diffusion of treatments, and the preparation material was distributed on a per treatment basis. Since the participants share the same social group, and they are interacting outside the research process as well, we cannot entirely rule out a cross-contamination between the groups.
- *Compensatory rivalry* is present if participants of a group put in extra effort when they have the impression that the treatment of another group might lead to better results than their own treatment. For example, participants of the LTL group might be aware that their assigned temporal property language is more difficult than PSP. We tried to mitigate the risk of compensatory rivalry by communicating that while there might be differences in difficulty, that would be considered in the grading process.
- *Demoralization* could occur if a participant is assigned to a specific group that she/he does not want to be part of. We did not observe any signs of demoralization such as increased dropout rates or complaints regarding group allocation.
- *Experimenter bias* refers to undesired effects on the dependent variables that are unintentionally introduced by the researcher. We tried to avoid such effects by designing the experiment in a way that limits any such chances. In particular, all participants worked on the same set of tasks (only the temporal property representation differs), and the results of the controlled experiment runs were processed automatically. The tasks used in the experiment were randomly generated, but there were similarities between the temporal properties used in some of the experiment tasks and those used in the examples discussed in the learning material. Such similarities might facilitate solving of related experiment tasks. To investigate this threat, we identified tasks that have similarities with the provided learning examples (cf. Table 9, Table 10, and Table 11). If there was a bias, such tasks should show a central tendency towards a relative high level of correctness while the remaining tasks should show a central tendency towards a relative low level of correctness. According to the acquired data, temporal properties with more predicates appear to be more difficult than those with less predicates, so we decided to normalize the measured correctness by the formula $correctness \times number\_of\_predicates \,/\, max\_predicates$ to enable a fair comparison between all tasks. We could not find any indication of bias introduced by those similarities in the gathered data. In particular, the number of possibly affected experiment tasks was almost balanced between the groups, and the measured correctness of possibly affected tasks was overall similar to those of the remaining tasks (cf. Table 9, Table 10, and Table 11). All approaches are presented by the same educational methods at a comparable level of detail to not introduce unnecessary bias into the experiment. A different choice of training material (e.g., formal semantics of LTL or the use of Structured English Grammar [14]) could have impacted the results. Also the design decision of using four instead of two truth value states (for a more fine-grained analysis of the understandability of a specification) might have had an impact on the results. Since however all groups had to cope with four runtime states, neither group was disadvantaged.

D.2 Threats to External Validity

The external validity is concerned with the generalizability of the results of our study. In the following, we discuss potential threats that hinder a generalization. There exist different types of generalizations that must be considered:

TABLE 9
Evaluation of the impact of similarities between experiment tasks and training examples on correctness in the first experiment run (* indicates similarity)

Pattern	# Predicates	PSP	EPL	LTL
Absence_AfterUntil	3	63.89	38.89 *	31.94
Absence_Between	3	78.33	71.67	52.50 *
Existence_AfterUntil	3	52.08	61.46 *	34.38 *
Existence_Between	3	81.25	54.17 *	39.58
Precedence_After	3	63.33	65.00 *	25.83
Response_After	3	52.78 *	50.00	54.17 *
Precedence_AfterUntil	4	75.00	40.83	36.67
Precedence_Between	4	77.08 *	45.83	16.67 *
Response_AfterUntil	4	40.00 *	36.67	25.00
Response_Between	4	56.67 *	42.50	20.83 *
Tasks similar to examples mean (not normalized)		56.63	54.88	35.71
Remaining tasks mean (not normalized)		68.98	47.92	31.80
Tasks similar to examples mean (normalized)		53.33	41.16	28.66
Remaining tasks mean (normalized)		54.86	42.85	26.94

TABLE 10
Evaluation of the impact of similarities between experiment tasks and training examples on correctness in the second experiment run - DSE (* indicates similarity)

Pattern	# Predicates	PSP	EPL	LTL
Absence_AfterUntil	3	69.44	64.29 *	40.32
Absence_Between	3	87.65	54.76	25.81 *
Existence_Between	3	77.78	63.10 *	46.24
Precedence_After	3	78.70	61.61 *	24.19
Response_After	3	74.81 *	60.71	54.84 *
Precedence_AfterUntil	4	67.59	50.89	32.26
Precedence_Between	4	45.68 *	54.76	19.35 *
Response_AfterUntil	4	56.30 *	45.00	18.71
Response_Between	4	77.04 *	29.29	30.32 *
Tasks similar to examples mean (not normalized)		63.46	63.00	32.58
Remaining tasks mean (not normalized)		76.23	49.24	32.34
Tasks similar to examples mean (normalized)		58.78	47.25	27.54
Remaining tasks mean (normalized)		60.55	44.42	26.81

TABLE 11
Evaluation of the impact of similarities between experiment tasks and training examples on correctness in the second experiment run - ASE (* indicates similarity)

Pattern	# Predicates	PSP	EPL	LTL
Absence_AfterUntil	3	61.76	63.24 *	40.63
Absence_Between	3	88.24	47.06	25.00 *
Existence_Between	3	88.24	58.82 *	54.17
Precedence_After	3	80.88	61.76 *	43.75
Response_After	3	81.18 *	60.00	47.50 *
Precedence_AfterUntil	4	72.06	47.06	46.88
Precedence_Between	4	50.98 *	52.94	18.75 *
Response_AfterUntil	4	51.76 *	49.41	16.25
Response_Between	4	76.47 *	49.41	35.00 *
Tasks similar to examples mean (not normalized)		65.10	61.27	31.56
Remaining tasks mean (not normalized)		78.24	50.98	40.34
Tasks similar to examples mean (normalized)		60.02	45.96	27.03
Remaining tasks mean (normalized)		62.28	46.52	33.41

- *Generalizations across populations*: By statistical inference, we try to make generalizations from the sample to the immediate population. The study considers two populations, namely computer science students that enrolled in the course DSE as proxies for novice to moderately advanced software architects, designers or developers, as well as computer science students that enrolled in the course ASE as proxies for moderately advanced software architects, designers or developers. The results of our study show similar results for both populations, but it is unclear to what extent these results are generalizable to different or broader populations. Therefore, we do not intend to claim generalizability without further empirical evidence. For example, it might be plausible that people working in the software industry with many years of experience or business administrators perform similarly, but the given study can neither support nor reject such claims.
- *Generalizations across treatments*: Since the treatments are equivalent to specific temporal property representations, treatment variations are inherently impossible.
- *Generalizations across settings/contexts*: The participants of this study are students who enrolled computer science courses at the University of Vienna, Austria. Apparently, a majority of the students are Austrian citizens, but there is a large presence of foreign students as well. Surely, it would be interesting to repeat the experiment in different settings/context to evaluate the generalizability in that regard. For example, the majority of the participants are non-native English speakers, which could be an obstacle for understanding the preparation material or task descriptions, so repeating the experiment with native speakers might lead to different (presumably better) results.
- *Generalizations across time*: We performed the experiment at two points in time (one year apart) with similar results. Especially, master students are rather heterogeneous group as they often come from other countries and faculties. This heterogeneity might explain the differences in previous experience with formal logic of master students in ASE between the first and second experiment run. Students in DSE appear to be more homogeneous, maybe because receiving training in formal logic is part of the bachelor program in computer science at the University of Vienna. In general, it is hard to predict whether the results of this study hold over time. For example, if teaching of LTL or EPL is intensified, then the students would bring in more LTL-related or EPL-related expertise, which likely has an impact on the results of the controlled experiment.

D.3 Threats to Construct Validity

There are potential threats to the validity of the construct that must be discussed:

- *Inexact definition & Construct confounding*: This study considers the construct *understandability* that is mea-

sured by the variables *correctness* and *response time*. To our best knowledge, this construct is exact and adequate. Several existing studies that evaluate different representations (e.g., domain specific languages) use this construct and its variables (cf. Feigenspan et al. [15] and Hoisl et al. [16]).

- *Mono-operation bias*: In this study, the independent variable is the temporal property language. Currently, we do not differentiate this construct any further. For example, the tasks of the experiment are based on a representative set of temporal property patterns with different numbers of propositional variables, but we do not perform further investigations on the basis of the number of propositional variables. Such finer-grained analyses are tempting, but a much larger number of tasks and/or answer choices would be necessary in order to be able to perform meaningful statistical analyses, and increasing the number of tasks and/or answer choices would likely result in experimental fatigue due to prolonged experiment sessions.
- *Mono-method bias*: To measure the correctness of answers, the evaluation by an automated method appears to be the most accurate measure as it does not suffer from experimenter bias or instrumental bias. For organizational reasons, keeping time records was the personal responsibility of each participant. Certainly, this leaves room for measuring errors, and an alternative measuring method (e.g., video records with timestamps or performing the experiment with an online tool that handles record keeping) would reduce the threat to construct validity. Participants who made obvious errors in their time records are not considered in this study (cf. Section A.1).
- *Reducing levels of measurements*: Both the correctness and response time are continuous variables. That is, the levels of measurements are not reduced.
- *Treatment-sensitive factorial structure*: In some empirical studies it might be the case that a treatment sensitizes the participant to develop a different view on the construct (e.g., differentiation between different types of stress). Since we did not ask questions regarding the subjective level of understandability of temporal property specifications in the controlled experiment runs, but tried to measure the actual level of understandability objectively, this threat is considered to be irrelevant.
  The survey questions asked in addition to the controlled experiment in the first experiment run are concerned with subjective preference rankings and subjective thoughts on practical applicability of the temporal property languages (or the lack of the same), so they are neither meant nor used to measure the understandability construct in this study.

D.4 Threats to Content Validity

Content validity is concerned with the relevance and representativeness of the elements of a study for the construct that is measured:

- *Relevance*: All tasks are based on the Property Specification Patterns (cf. Dwyer et al. [17]), which is a set of commonly occurring temporal property patterns. Thus, we claim that the contents of the experiment are highly relevant for measuring the understandability of temporal property representations. However, using the patterns as basis for our tasks might be a threat to validity for measuring the understandability of LTL and EPL, because the expressiveness of these approaches goes far beyond the pattern-based approach, which is limited to a set of patterns. In that regard, for future work, it would be interesting to design an experiment that focuses on LTL and EPL with tasks that are not based on patterns. In the context of the presented study, it was necessary to base the tasks on patterns, otherwise it would not have been possible to include PSP in the study.
- *Representativeness*: A representative subset of existing Property Specification Patterns was used for the tasks of the experiment. To reduce chances for experimental fatigue, we did not include all of the available patterns, but we selected the most commonly used patterns according to Dwyer et al. [17]. A survey by Bianculli et al. [18] based on 104 scientific case studies reproduced the results of the survey in [17] even 13 years after the original study took place. In particular, the *Response Chain*, *Precedence Chain*, *Constrained Chain* and *Bounded Existence* patterns are omitted, because they are rarely used. The study by Bianculli et al. [18] investigated the PSP used in a set of industrial service-based applications. Interestingly, the patterns found in the requirement specifications of 100 randomly selected service interfaces were rather concerned with non-functional requirements like the maximum number of events in a certain time interval within a certain time window than the qualitative order or existence/absence of events. That is, patterns used in practice might be different from those in scientific studies. Please note, however, that the generalizability of these results is rather limited as the results might only apply to service-oriented computing in that specific company.

D.5 Threats to Conclusion Validity

Retaining outliers might be a threat to conclusion validity. However, all outliers appear to be valid measurements, so deleting them would pose a threat to conclusion validity as well. We performed a thorough evaluation of the model assumptions of all relevant statistical tests and selected the test with the greatest statistical power. That course of action is considered to be extremely beneficial to the conclusion validity of this study.

REFERENCES

[1] C. Czepa and U. Zdun, "On the Understandability of Temporal Properties Formalized in Linear Temporal Logic, Property Specification Patterns and Event Processing Language [Data set]," http://doi.org/10.5281/zenodo.891007, 2017.

[2] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust statistical methods for empirical software engineering," *Empirical Software Engineering*, pp. 1–52, 2016.

[3] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychological Bulletin*, vol. 114, pp. 494–509, 1993.
[4] J. J. Rogmann, "Ordinal dominance statistics (orddom): An r project for statistical computing package to compute ordinal, nonparametric alternatives to mean comparison (version 3.1)," Available online from the CRAN website http://cran.r-project.org/, 2013.
[5] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
[6] A. R. da Silva, G. Malafaia, and I. P. P. de Menezes, "biotools: an r function to predict spatial gene diversity via an individual-based approach," *Genetics and Molecular Research*, vol. 16, p. gmr16029655, 2017.
[7] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 2nd ed. Thousand Oaks CA: Sage, 2011. [Online]. Available: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
[8] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. [Online]. Available: http://ggplot2.org
[9] Slawomir Jarek, "mvnormtest: Normality test for multivariate variables," https://CRAN.R-project.org/package=mvnormtest, 2012, last accessed: May 30, 2018.
[10] Peter Filzmoser and Moritz Gschwandtner, "mvoutlier: Multivariate Outlier Detection Based on Robust Methods," https://CRAN.R-project.org/package=mvoutlier, 2017, last accessed: May 30, 2018.
[11] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois, 2017, r package version 1.7.5. [Online]. Available: https://CRAN.R-project.org/package=psych
[12] B. Naimi, N. a.s. Hamm, T. A. Groen, A. K. Skidmore, and A. G. Toxopeus, "Where is positional uncertainty a problem for species distribution modelling," *Ecography*, vol. 37, pp. 191–203, 2014.
[13] H.-F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277–1288, 2005, pMID: 16204405. [Online]. Available: http://dx.doi.org/10.1177/1049732305276687
[14] M. Autili, L. Grunske, M. Lumpe, P. Pelliccione, and A. Tang, "Aligning qualitative, real-time, and probabilistic property specification patterns using a structured english grammar," *IEEE Transactions on Software Engineering*, vol. 41, no. 7, pp. 620–638, July 2015.
[15] J. Feigenspan, C. Kästner, S. Apel, J. Liebig, M. Schulze, R. Dachselt, M. Papendieck, T. Leich, and G. Saake, "Do background colors improve program comprehension in the #ifdef hell?" *Empirical Software Engineering*, vol. 18, no. 4, pp. 699–745, 2013.
[16] B. Hoisl, S. Sobernig, and M. Strembeck, "Comparing three notations for defining scenario-based model tests: A controlled experiment," in *QUATIC'14*, Sept 2014, pp. 95–104.
[17] M. B. Dwyer, G. S. Avrunin, and J. C. Corbett, "Patterns in property specifications for finite-state verification," in *Proceedings of the 21st International Conference on Software Engineering*, ser. ICSE '99. New York, NY, USA: ACM, 1999, pp. 411–420. [Online]. Available: http://doi.acm.org/10.1145/302405.302672
[18] D. Bianculli, C. Ghezzi, C. Pautasso, and P. Senti, "Specification patterns from research to industry: A case study in service-based applications," in *2012 34th International Conference on Software Engineering (ICSE)*, June 2012, pp. 968–976.