CHAPTER 5

Evaluating the Effects of Adaptation based on Learners' Traits

In Chapter 4, we have described the design and results of the empirical evaluation on the *media* aspect, which compared the effectiveness of using different media representations on the task of spatial ability enhancement.

In this chapter, we focus on evaluating another aspect of the "M&M" concern: from the viewpoint of the method aspect, to evaluate the effects of using a different mechanism to select learning materials to learners. It is specifically intended to investigate the value of adaptivity in terms of learners' traits, including spatial ability and learning styles.

Some studies have been endeavored on comparing the effects of with- and without- adaptivity in learning environments. Most of them focus on evaluating the effectiveness of adaptivity in terms of learners' *knowledge*. Or in other words, these studies were undertaken to compare the difference between a system that can adapt to learners' knowledge and another system that cannot do so [17]. Comparatively, little research has been undertaken on developing AH systems based on learners' potential ability and learning/cognitive styles [62][66]. Among them, empirical evaluation on the effects of adaptation based on learning styles is even just at the beginning. Here we present our experimental design and results on the issue of adaptation regarding learners' traits.

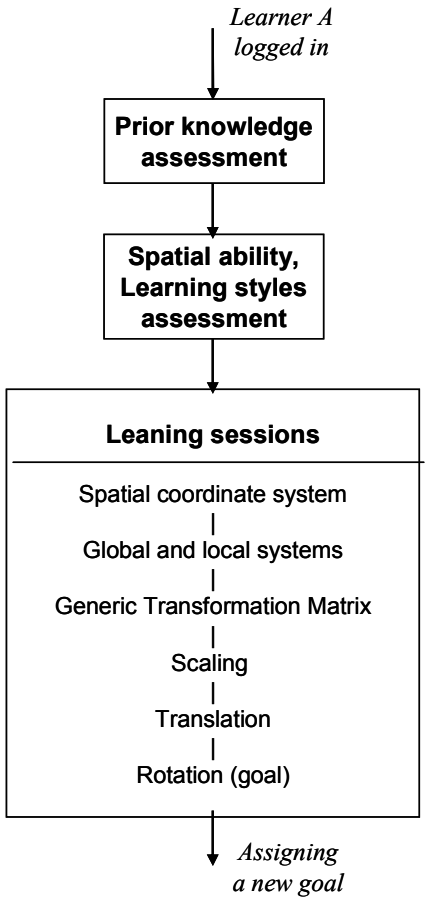

Figure 5. 1: Major events that learners will meet by setting the concept of *Rotation around single axis* as the learning goal.

5.1 Walkthrough on how CooTutor works

We have described the mechanisms of concept sequencing and material selection in Chapter 3. Before describing the empirical evaluation, here we present how the adaptive mechanisms may influence the presentation, specifically how learners with different traits, including spatial ability and learning styles, would be tutored adaptively in CooTutor. A simple cognitive walkthrough [58] is presented here to demonstrate how the system actually works.

Assumes that learner *A* has logged in CooTutor. Meanwhile, *A*'s learning goal is to learn the concept of *Rotation around single axis*. The major events that she/he will meet are illustrated in Figure 5.1. Assume that the assessment of prior knowledge indicates that *A* has known concepts of *Fundamental matrix* and *Matrix multiplication*. The system thus could ar-

Table 5. 1: Learning materials with feature values and similarity measures.
Rows are grayed if that learning material is not recommended. ($S_{threshold}$ = 0.7)

Concept	ID	Main_rep.	Abstractness	Activity_type	Similarity	Recommended
Spatial Coordinate System	*#27*	0.2	0.2	0.2	0.90	T
	#25	0.2	0.8	0.2	0.71	T
	#26	0.2	0.9	0.2	0.65	F
Global and Local Sys.	*#31*	0.4	0.3	0.2	0.90	T
	#30	0.4	0.3	0.2	0.90	T
Generic Trans.	*#29*	0.2	0.7	0.2	0.77	T
Scaling	*#53*	0.2	0.8	0.2	0.71	T
	#33	0.8	0.2	0.9	0.68	F
Translation	*#52*	0.2	0.8	0.2	0.71	T
	#32	0.8	0.2	0.9	0.68	F
Rotation around Single Axis	*#36*	0.2	0.3	0.2	0.91	T
	#35	0.2	0.4	0.2	0.90	T
	#37	0.8	0.2	0.9	0.68	F
	#34	0.2	0.9	0.2	0.65	F

range a learning plan consisting of six concepts by the algorithm of concept sequencing, as shown in Figure 5.1. Note that concepts known by the learner will not be arranged into the plan. Assume that learner *A*'s spatial ability score is assessed via the PVRT test as a normalized value (0-1) of 0.8 (i.e., a learner with sufficient spatial reasoning skill); sensing/intuitive learning style is assessed by the learning style questionnaire as 0.2 (i.e., a sensing-apt learner); active/reflective learning style is assessed as 0.5 (i.e., a balanced learner respect to this dimension of learning style). Then the following query is generated by the stereotype generator mentioned in section 3.3.4:

Q = <*is_2D*, *is_3D*, *is_concrete*, i*s_abstract*, *is_lecture*, *is_experiment*, *level_of_details*>
= (0.7, 0.3, 0.7, 0.3, 0.5, 0.5, 0.5)

By employing the mechanism of material selection, the similarity measures of learning materials are computed. And the order of recommendation is ranked as well. Table 5.1 illus-

trates all learning materials in the content repository that relate to learning these concepts. Assume that the threshold is set as 0.7. It could be observed from Table 5.1 that besides the order of presentation is ranked according to the similarity score, some of the learning materials are not recommended as well. For example, the *#26* learning material is not recommended. Its features <*Main_representation*, *Abstractness*, *Activity_type*> is (0.2, 0.9, 0.2), and could be transformed as the feature vector:

$$M = \langle is\_2D, is\_3D, is\_concrete, is\_abstract, is\_lecture, is\_experiment, level\_of\_details \rangle$$

$$= (0.8, 0.2, 0.1, 0.9, 0.8, 0.2, 0.5)$$

The similarity measure is computed as 0.65, which is lower than the threshold, so the learning material is not recommended.

CooTutor also employed the adaptive method—adaptive navigation support [10][14] in order to let learners reflect and grasp their own learning progress. Figure 5.2 illustrates the

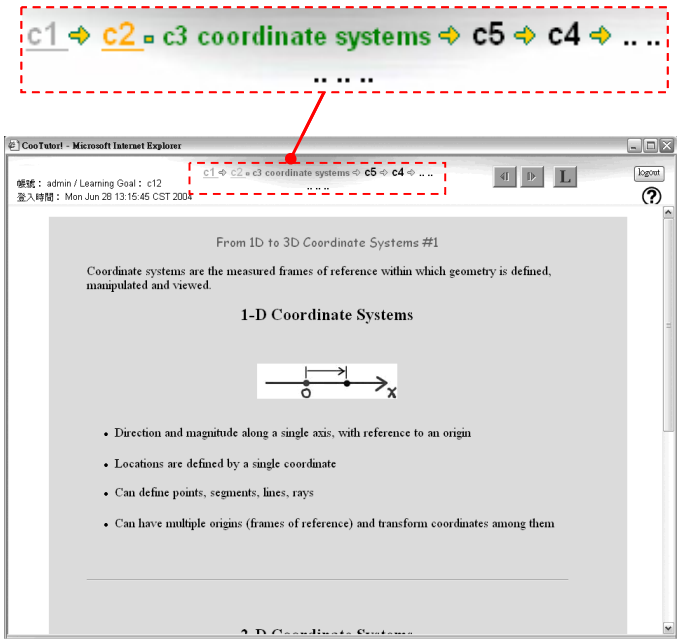

Figure 5. 2: Adaptive Navigation Support offered by CooTutor.
Links are colored for the purpose of guidance.

navigation bar of CooTutor, which is the enlarged part in the figure. In the navigation bar, clickable links of "concepts" are annotated and colored with metaphor as the guidance. Grayed links, like "C1" in Figure 5.2, imply that the concepts have been known by the learner. The links colored orange, like "C2" in the figure, mean that the learner has learned this concept but does not know well. Whether a concept is learned well is determined by a simple quiz during the learning sessions. The links or descriptions colored green, mean that the learner is now learning the concept, or that concept is ready to be learned. For learner *A* in this case, several fundamental concepts would be initially annotated as gray color accommodating to the status held by the student model.

CooTutor aims to offer guidance, but not restriction to learners' browsing. The simplest navigational method is to directly click on "NEXT" or "BACK" buttons shown in the user interface to navigate all of recommended learning materials. The order of presentation reflects the order of recommendation shown in Table 5.1. Once the learner finished viewing recommended materials of a specific concept, materials of next concept would be fetched and presented subsequently. Also, the status (i.e., color) of the navigation bar as shown in Figure 5.2 will be updated. However, un-recommended materials will not be arranged into the presentation queue.

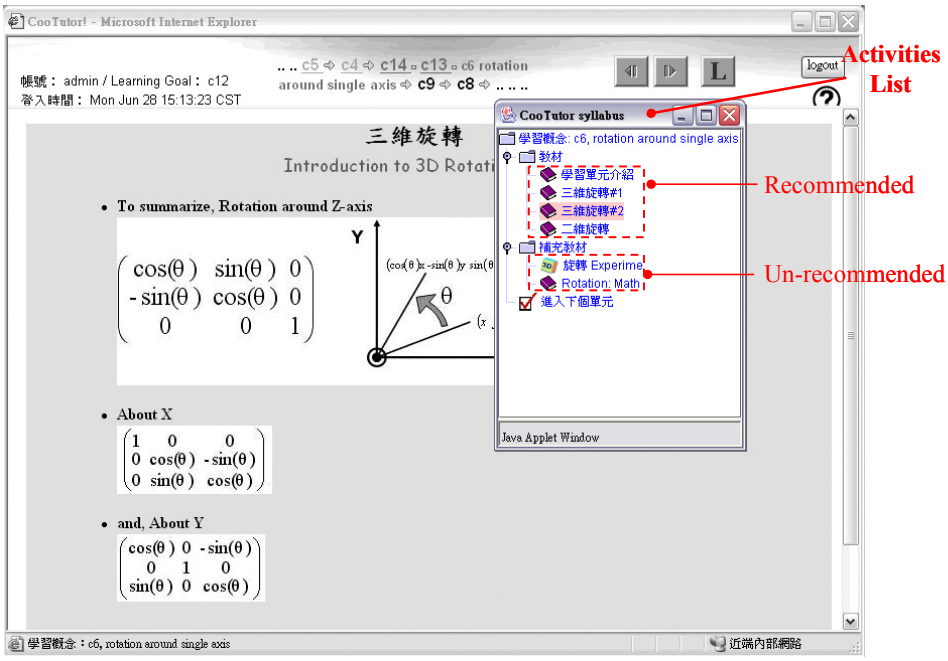

Figure 5. 3: Activity list presented to the learner showing the recommendation.

The learner can also choose to browse the hyperspace with jumps within certain scope. Evidently, by the design of adaptive navigation support, learners are not eligible to browse learning materials associated with concepts that are not ready to be learned, unless the learner has acquired prerequisite concepts. However, learners are allowed to review concepts they have known and those they have learned but not very well. By visiting each concept, CooTutor would prepare an activity list for the learner to navigate without a specific sequence. Figure 5.3 shows the activity list presented to the learner for learning the concept of "Rotation around single axis". By adopting the activity list, learners can even browse contents that are not recommended stereotypically by the system.

A simple walkthrough on how CooTutor works has been presented. Several adaptive methods have been realized by CooTutor's architecture. These methods mainly include material sequencing regarding style-matching between learning and pedagogical styles as well as adaptive navigation support based on learning status (i.e., learners' knowledge or performance). Next, the empirical evaluation focused on adaptation based on *learners' traits* will be

presented.

5.2 Design of the experiment

This experiment aims to probe the effectiveness of traits-based adaptation in CooTutor. Four different versions of CooTutor with different methods on material selection were used in this experiment. Similar to the experiment described in Chapter 4, this experiment adopted a pre-test/post-test comparison-group experimental design. These four versions of system differ from each other on the strategy of material selection.

The experiment was held in June 2004 at National Chengchi University (NCCU). Totally 31 graduate-level participants majored in Computer Science or Information Systems (master program) from NCCU have attended the experiment. All of them have learned fundamental linear algebra and computer graphics. They were grouped as four groups, while each group was assigned to use one version of CooTutor. The process of grouping is double-blinded but not totally random. That is, all participants did not know what version of the system they used. And the experimenter did not know the participants' pre-test scores of spatial ability or domain achievement test and thus could not purposely prefer any group to others.

The whole duration of the experiment lasted for three weeks. Participants are asked to log in the system, take the pre-tests, view all learning materials, and finally be tested by post-tests. However, we intend to make the experiment conforming to the characteristic of Web-based learning as mentioned in Chapter 4, the characteristic of self-paced learning. Therefore, participants were not enforced to operate the system at a specific time and fixed duration, such as an hour. They were only informed by the experimenter that they can log into the system at any moment they want before a specific due date. All of the four groups are assigned the same learning goal—the concept of "*Gimbal Lock*" in SGT. And the same learning plan with identical concept sequence consisted of 14 concepts is used by all groups. In other words, the factor of concept sequencing is under controlled between all groups. All of the four groups would all receive partial degree of adaptation in terms of knowledge, including concept sequencing and adaptive navigation support. The treatment of the experiment is thus with or without adaptation in terms of learning styles. Figure 5.4 depicts the process of the experiment.

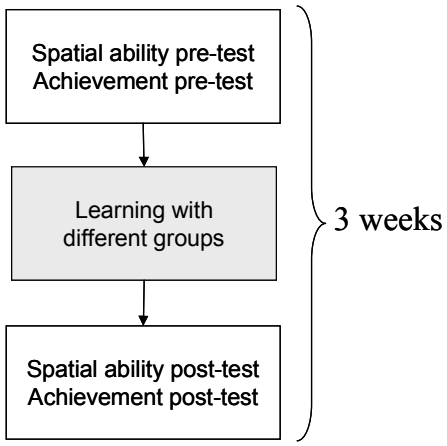

Figure 5. 4: The design of experiment

Table 5. 2: Four versions of CooTutor system compared in the experiment

Group	Num. of Participants	Strategy of Material Selection	Adaptive Ranking?	Size of hyperspace (# of materials)
LS Style-matching group	10	Traits matching	YES	Varies
PreAuthor Pre-authored group	12	Pre-authored and selected manually	NO	16
NoFilter No-filtering group	4	No material selection	NO	33
MisLS Style-mismatching group	5	Learning style mismatching	Mismatch	Varies

Table 5.2 summarizes how these four versions of CooTutor differ. Note that the number of participants in each group differs. Especially the last two groups shown in the table only have around 5 participants each. This is because our main interest is actually upon the first two groups (i.e., the *LS* and *PreAuthor* groups), but the effect of the last group is also suspected. In order to address both issues, we choose to assign a large portion of participants into the first two groups, but keep a small portion of them in the last two for references.

Among these four versions, *LS* is the version that employs the mechanism of adaptive material selection proposed in section 3.3.4. And the score of threshold of recommendation was set as 0.6. Therefore, learning materials would be selected and ranked adaptively based on participants' traits. Note that since learners' spatial ability and learning styles would vary, so the size of hyperspace would vary by the use of adaptive material selection as well. In other words some inappropriate materials would be probably filtered to a particular learner.

PreAuthor is the version that does not use such a adaptive mechanism, but ask a domain expert of SGT to select a fixed set of learning materials a priori. This version could be thought as the group without traits-based adaptivity. Totally 16 learning materials were pre-selected for learning these 14 concepts. The third group, *NoFilter*, is the group that offers the participants all available materials stored in the content repository which now holds 33 learning materials. That is, no filtering or selection would be done to reduce the hyperspace.

The last version, *MisLS* is the version that designed to probe *what if learning materials with totally inverse pedagogical styles regarding learners' learning styles (i.e.,styles-mismatching) were presented to the learner*. Note that in section 3.3.5, we have described the method of theory refinement based on learners' feedback. Consequently, some elements of learners' query vector Q are inherently variants. For example, elements of *is_2D*, *is_3D*, *is_concrete* and *is_abstract* in the query vector are designed to be refinable. For properly controlling experimental factors, we choose to only mismatch the element *is_lecture* and *is_experiment* of Q. The strategy is to exchange the values of *is_lecture* and *is_experiment*. Clearly, the premise of this experimental treatment is that the learner must have an extreme score on this dimension of learning style (i.e., the active/reflective learning style) and thus "mismatching" of learning style can be possible. There are 5 participants with extreme score on this learning style dispatched to this group. This is why we have mentioned that the grouping is not totally random.

Some decisions have been undertaken to group participants strategically. This is also why a pre-test on spatial ability and achievement was performed to all participants. This will help us to compare the effects and interpret the results of different groups with fair.

5.3 Measuring instrument

There were three types of score measured in this experiment. They are *learners' spatial ability*, *achievement on the topic of spatial geometric transformation (SGT)*, and *learners' attitudes on CooTutor*. For measuring spatial ability, the Web-based PVRT mentioned in Chapter 4 was employed again. For measuring learners' achievement on SGT, a self-compiled achievement test consisting of 7 items was authored and integrated into CooTutor. And for measuring learners' attitudes on the system, an attitude questionnaire consisting of 15 single-choice questions was used in the task.

We have described details of the Web-based PVRT in Chapter 4, so we do not replicate the description here. The estimated reliability by using KR-20 of this experiment is reported here: 0.69 for the pre-test and 0.75 for the post-test. It is worth noting that the KR-20 coefficient is recognized acceptable but not as high as the one derived in the previous experiment of Chapter 4. For the fact that reliability coefficient is a characteristic of data, it is reasonable to detect such a difference (Huck 2000, p. 98) [40]. Specifically this experiment is even closer to real scenario of Web-based learning with a long duration of intervention.

The 7-item SGT achievement test was compiled and employed on both pre- and post-test. Appendix A shows some sample items of this instrument. The estimated KR-20 coefficient is 0.59 for the pre-test and 0.28 for the post-test. Note that the coefficient reveals low internal inconsistency from the viewpoint of classical test theory [32]. It is suspected that two main factors are subject to the scenario. *First*, the number of items of this test is rather small. It is naturally difficult to achieve high reliability for a test with only few items. *Second*, the homogeneity between participants is high. As the result that will be described in section 5.5 reveals, all participants performed quite well on the achievement post-test. That is, amount of participants has been very close to the limitation of measurement of this instrument. Since the underlying computation of reliability coefficient relates to the distribution of measuring results a lot [32]. When the distribution is highly skewed like this case, it is likely that the coef-

ficient would be influenced to be low. It is interesting to note that at one hand the low reliability coefficient on post-test seems suggested that the result is not reliable. But on the other hand, by taking the global observation on results of both pre- and post-test, it is very likely that learning with CooTutor has substantially changed the distribution of the participants toward the high score area. Our observation suggest that, the instrument itself is *probably not* that un-reliable (KR-20 0.59 on the pre-test), but the improvement of SGT understanding has contributed side-effects to computing the reliability coefficient of post-test.

The 15-item attitudes questionnaire is used to assess learners' attitudes toward CooTutor. Question items of the questionnaire are shown in Appendix B. This questionnaire adopts the 6 point Likert-type scoring method. For each question item, there are six response options extending from "strongly agree" to "strongly disagree" and are scored from 6 to 1 correspondingly. The reliability coefficient by using the method of Cronbach's alpha [32][40] is estimated as 0.96. The result of this questionnaire is quite reliable. In order to probe participants' multi-dimensional attitudes toward the system, partial items were specifically grouped and analyzed separately for assessing a specific type of attitude that we are interested in. Three sub-groups were identified. They are *system's guidance*, *system's recommendation* and *learners' learning engagement (or interestingness of SGT)*. Appendix B also shows the scenario.

5.4 Data analysis

Since the design of experiment is quite similar to the previous experiment presented in Chapter 4, we can still apply the ANCOVA method here by using pre-test as the covariate. However, note that there are totally four groups involved in this experiment, and the number of participants (i.e., degree of freedom of statistics) differs between them. For the groups we are mainly interested in—*LS* and *PreAuthor* groups, each group was assigned around 10 participants. For groups intended to be as references, each group was assigned around 5 participants. Under this scenario of uneven number of participants, it is *not* tenable to let all groups involving in ANOVA or ANCAOVA analysis for the concern of unbalanced degree of freedom among different populations [4]. Thus we only compare the results of *LS* group and *PreAuthor* group by using ANCOVA, while for other two groups, we reveal the results by reporting descriptive statistics and effect size.

5.5 Experimental Results

Table 5.3, 5.4 and 5.5 present the result of *PVRT test*, *achievement test* and *attitude questionnaire* respectively. For tests consisting of both pre- and post-tests, i.e., the PVRT test and achievement test, a paired 2-tailed *t* test was performed to compare the difference of means between post- and pre- test. Meanwhile, the effect size is specifically computed by using Cohen's *d* coefficient [18]. The formula of Cohen's *d* is:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} \tag{5.1}$$

where μ_1 and μ_2 are the means of populations in comparison with each other. σ_{pooled} is the pooled standard deviation combining the variance from two populations (i.e., post- vs. pre- tests). In most statistical analysis, it is usually assumed that two populations have the same variance. However, it seems unlikely that the distribution of post-test scores would have the same variance with the pre-test in this case. For example, in Table 5.3, both *LS* and *Pre-Author* groups derive a smaller standard deviation on the post-test (the SD is around 3 for the pre-test, but 2 for the post-test). Hence, to calculate the pooled standard deviation is deemed essential in the task of estimating effect size. That is:

$$\sigma_{pooled} = \sqrt{(\sigma_1^2 + \sigma_2^2)/2} \tag{5.2}$$

, σ_1 and σ_2 are the standard deviations of the two distributions to be compared.

Since Cohen's *d* coefficient (and other type of effect size measure) is a standardized score, as we have mentioned in Chapter 4, some criterion is needed to judge and conceptualize the result. From the literature, researchers have suggested such a criterion used here. That is, for Cohen's *d* coefficient, 0.2 is a small effect size; 0.5 implies medium size; 0.8 and above indicates a large effect size [4][18]. By using the viewpoint of effect size, we describe the result of each part of experiment respectively.

Table 5. 3: Statistics of participants' pre- and post- PVRT scores along with Cohen's d coefficient [18] for indicating the effect size.

	Pre_PVRT		Post_PVRT		Post vs. Pre *t*-test, p	Effect size, *d*
	Mean	SD	Mean	SD		
LS (n=10)	15.600	3.026	16.600	2.119	0.437	0.383
PreAuthor (n=12)	13.250	3.223	16.167	2.125	0.013**	1.069‡
NoFilter (n=4)	17.000	2.160	18.250	2.217	0.194	0.570†
MisLS (n=5)	16.600	1.949	13.200	5.805	0.252	-0.785Δ
Overall	15.032	3.136	16.097	3.177	0.173	0.337

*p<0.1 **p<0.05 Effect size: ‡large, †medium, Δ negatively large

Table 5. 4: Statistics of participants' pre- and post- achievement test scores.

	Pre_Achievement		Post_Achievement		Post vs. Pre *t*-test, p	Effect size, *d*
	Mean	SD	Mean	SD		
LS (n=10)	3.900	1.853	4.900	1.524	0.052*	0.589†
PreAuthor (n=12)	4.667	1.723	5.333	1.231	0.166	0.445†
NoFilter (n=4)	4.000	1.633	5.250	1.258	0.015**	0.857‡
MisLS (n=5)	4.400	1.817	4.600	0.894	0.749	0.140♦
Overall	4.290	1.716	5.065	1.263	0.003***	0.514†

*p<0.1 **p<0.05 ***p<0.01 Effect size: ‡large, †medium, ♦small

Table 5. 5: Statistics of participants' response on the attitude questionnaire.

	Guidance		Recommendation		Engagement		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
LS (n=10)	3.833	1.189	3.350	1.226	3.700	1.311	3.647	1.145
PreAuthor (n=12)	4.056	0.930	3.542	1.157	4.083	0.660	4.006	0.747
NoFilter (n=4)	4.250	1.229	3.500	0.707	3.938	1.008	4.050	0.900
MisLS (n=5)	4.333	1.000	4.300	1.204	5.100	0.978	4.773	0.992
Overall (n=31)	4.054	1.030	3.597	1.136	4.105	1.062	4.019	0.975
LS vs. *PreAuthor* t-test, p	0.628 (*d*=0.2)		0.710 (*d*=0.16)		0.384 (*d*=0.36)		0.387 (*d*=0.37)	

5.5.1 Result of spatial ability enhancement

In Chapter 4, we have investigated the effects of different media representations on enhancing learners' spatial ability. We are now interested in a similar problem—to probe if different strategies of material selection would influence the effects of spatial ability enhancement. An ANCOVA analysis has been conducted upon the data of *LS* and *PreAuthor*. The result, $F(1,19)=0.194$, $p=0.664$, is *not* statistically significant. And for the effect size of this comparison, $\eta^2 = 0.01$ indicates that only small effect size existed.

Back to Table 5.3, we can take a global view on all of the statistics. It is worth noting that the *PreAuthor* group reveals the best performance among all groups on spatial ability enhancement. The result of paired t-test reveals statistically significant ($p=0.013 < 0.05$), and the effect size is quite large ($d=1.069$). Though the *LS* group did not reveal strong effectiveness on this task, but on the other hand the *MisLS* group performs quite worse on the post-test, the effect size, $d=-0.785$, is very large on the *inverse* (i.e., negative) direction. There is no similar scenario happened to other groups.

5.5.2 Result of SGT achievement

The ANCOVA analysis comparing *LS* and *PreAuthor* shows no statistical significance as well. The result is: $F(1,19)=0.034$, $p=0.856$. The effect size of the difference, $\eta^2 = 0.002$ is quite small. That is, it could be inferred that these two groups performed almost equally well on the task of enhancing SGT achievement.

From Table 5.4, the *LS* group seems still a little bit better than the *PreAuthor* group. From the view of effect size, the *d* coefficient is 0.589 for the *LS* group which is a medium effect size, and 0.445 for the *PreAuthor* group which is very close to a medium one. The best group is the *NoFilter* group, the result of paired *t*-test comparing the means of post- and pre- tests indicates statistical significance ($p=0.015<0.05$). Its effect size, $d=0.857$ is a large one. At last, for the *MisLS* group, though there shows an increasing scenario of the post-test comparing to the pre-test. However, the gain effect is very small. The effect size $d=0.140$ is small and distant from the degree of improvement shown by other groups.

5.5.3 Learners' attitude on CooTutor

Table 5.5 shows the result of participants' attitude on CooTutor. Since the questionnaire adopts the 6 point Likert-type scoring method. That is, for each question item, 1 is the lowest score, and 6 is the highest one. Therefore, the middle score of each item is (1+6)/2 = 3.5. It can be observed that in Table 5.5, most results exceed the middle point indicating that participants have positive response on the system.

A two-tailed independent *t*-test is conducted upon the scores of *LS* and *PreAuthor* groups. For scores of guidance, recommendation, learning engagement and the overall score, no statistical significance is found. The effect size *d* of the comparison (i.e., *LS* vs. *PreAuthor*) on each category is all small. Tough it seems interesting that the *LS* group got somewhat lower scores than others on each category, but the effect size is small. It is difficult to judge if the underlying attitude of this group's participants is evidently different to other groups'.

Table 5. 6: Statistics of participants' usage of the system

Group	Login times (#)		Time spent each login (min)		Total time spent (min)	
	Mean	SD	Mean	SD	Mean	SD
LS	2.50	1.07	15.15	11.39	35.60	23.07
PreAuthor	2.58	1.44	12.65	9.62	32.69	21.72
NoFilter	2.75	1.71	15.09	9.10	41.50	18.02
MisLS	2.40	1.14	19.92	17.76	47.79	19.73

5.5.4 Learners' usage behavior

Participants' usage of the system is also concerned in this experiment as a complement to test scores. Table 5.6 shows the result. Note that raw data have been cleaned and filtered appropriately. For each login record, if the usage time is below one minute, then the record is abandoned. For all four groups, participants approximately visited the system 2.5 times. For the duration of each login, the result is approximately between 12 to 20 minutes. And for the total time spent on using the system, different groups differ a lot. For *LS* and *PreAuthor* groups, participants spent less time to use the system. While for *NoFilter* and *MisLS* groups, the participating time is larger than the previous two groups.

The interpretation of the result could be quite contradictory. The critical point is about *whether it is good or not that a user stayed long?* Especially when learning on the Web is actually self-directed and self-paced, if the system can attract learners to stay long, this is not a bad news. However, from another point of view, many studies in AH recognize that to stay shorter would be better [2]. The underlying logic is, if two groups of learners using with- and without- adaptivity systems respectively, no significant difference on achievement, but the two groups differ in using time, then this could be inferred that the group with less using time is more efficient than another. We acknowledge that AH is potentially for learners to learn efficiently because of adaptive guidance. Learners would not need to spend time on things they have known. However, such type of comparison and inference seems problematic. Specifically, what if the efficiency is due to learners' sloppily browsing behavior? Therefore, we intend not to judge the result shown in Table 5.6 strictly. Some theory has been proposed to address the meaning of such behavior patterns [43].

5.6 Discussion

The discussion starts from investigating the problem of questionnaire. Using attitude questionnaire to evaluate the system is a common and popular method been widely applied. In the field of Web-based learning (or instruction), this type of evaluation is convenient for both researcher and participants [78]. To author or answer a questionnaire is much easier than instruments like achievement tests or formally psychometric tests. However, it is inevitably that such type of measurement might be quite inaccurate and un-reliable as the example of drinking behavior illustrated by Underwood et al. in [73]. In this example, participants of that

ing behavior illustrated by Underwood et al. in [73]. In this example, participants of that experiment were asked to fill out the questionnaire indicating how frequently they drink alcoholic drinks. Besides, researchers also collected the data implicitly (i.e., participants were unaware) from surveying drink bottles appearing in participants' trash can. The example revealed that the data reported by participants are much lower than the fact. Yu et al. indicated that people may tend to "…(a) report what they believe the researchers expects to see, or (b) report what reflects positively on their own abilities. [78]" The lesson we learned from these cases is clearly that this type of self-reported data should be undertaken cautiously.

Relating it back to the result of this experiment, from Table 5.3 it is found that the *MisLS* group performs a rather degree of decrease on the PVRT post-test comparing to their high

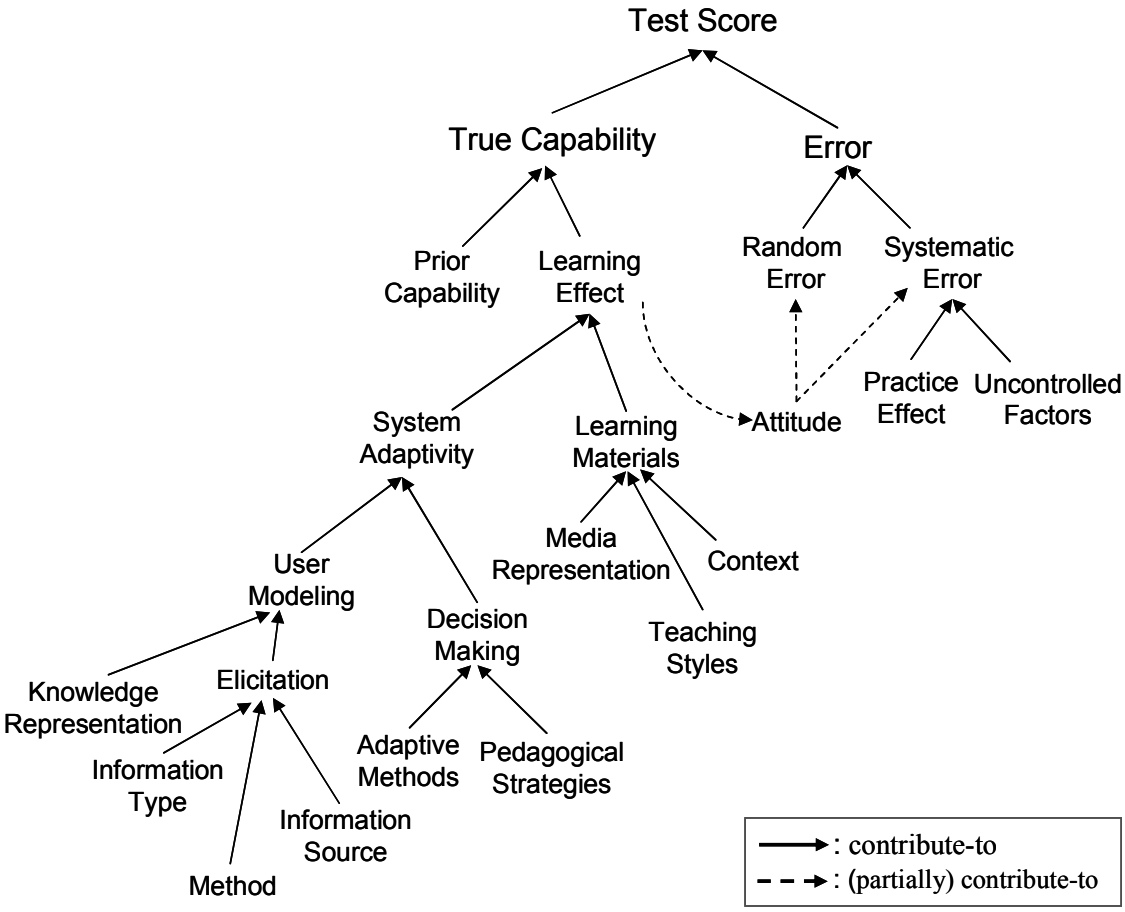

Figure 5. 5: Factors underlying the test scores of the experiment

scores on the pre-test. Clearly, this scenario cannot be interpreted as that the participants become 'stupid' when they did the post-test. It is suspected that this scenario is largely caused by the attitude underlying them, especially, the attitude of not willing to give their best effort on the post-test. If the deduction is true, then the questionnaire should reflect such situation ideally. However, the fact is that *MisLS*'s questionnaire score is quiet high. For the category of engagement, *MisLS* even shows the highest score among all the groups. It is suggested that attitude questionnaire is best to be used along with other instruments or data source (e.g., Web usage mining). To simply use self-reported data from questionnaire might be inappropriate to reflect learners' underlying attitude.

Another point to be elaborated here is about the validity of empirical evaluations. As we just mentioned, the underlying factors behind test scores confuse experimental psychologists a lot. In this case, the attitude questionnaire seems to derive inaccurate measurement on "attitude", while on the other side the difference of post- and pre- PVRT scores of *MisLS* seems to reflect that their attribute contribute to the scores. Less empirical study in the AH field has addressed this scenario formally. We are interested in asking *what factors are there underlying the evaluation of AH systems?* From the view of classical test theory, test scores derived (i.e., observed) by using instruments could be decomposed as [7][32]:

$$S_{observed} = S_{true} + E \tag{5.3}$$

where $S_{observed}$ is the observable score derived from a test, S_{true} is the unknown true score and E is the measurement error. By taking the notion of such decomposition, a meta-analysis of this experiment is depicted as a *score-contribution tree* shown in Figure 5.5. The "test score" shown in the figure refers to observable scores, such as scores of PVRT test or achievement test. The nodes of the tree refer to tasks, constructs or factors that are recognized contributing to the observable score. Two kinds of edge are shown here. Solid line means "contribute-to", such as the "error" would contribute to (i.e., influence) the "test score". Dashed line means uncertainty. We suspect that "attitude" could be random error or systematic error. That is, if after using some systems, learners' attitude would be largely skewed. Then this cannot be anymore just a random error, but should be a systematic one.

Based on this tree structure, it is clear that what researchers of the AH field want to evalu-

ate mostly lies on the bottom of the tree. In other words, there exist several layers between the factors we are of interest about and the observable measure. This scenario greatly challenges the validity of empirical evaluation of AH. Therefore, we suggest that besides statistical significance tests and the effect size, researchers should best report the latent factors of the experiment seen by themselves. Since most AH systems are proposed and built within the laboratory nowadays, for readers outside the laboratory it is difficult to probe if there were confounding factors or how serious it might be. A visual analysis such as Figure 5.5 depicted can also help practitioners within the laboratory (i.e., the AH field) to communicate clearly.

Summarizing the result revealed in this experiment, although adaptive material selection regarding learning styles does not outperform to other design, especially the version we intended to compare with, a set of learning materials selected by a human teacher. However, it is worthy noting that styles mismatching might yield negative effects on learning, specifically for those learners with extreme learning styles.

By this experiment, it is suggested that the mechanism of adaptive material selection is applicable, especially to prevent severe scenario of mismatching. On the other hand, according to Felder et al. [30], the best strategy to tackle learning styles may not be considering how to match the pedagogical styles to the learner, but about how to address each style evenly in the instruction. We also recognize that besides adaptive material selection it would be beneficial to consider adding adaptive facilitation to support learners learning from materials they do not like or prefer. For example, it is difficult to ask a teacher to teach mathematics without using mathematical descriptions, especially for those advanced topics. We recognize that tutoring learners how to learn would become a new challenging and topic for AH systems.