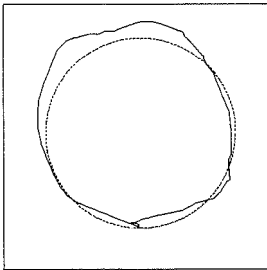
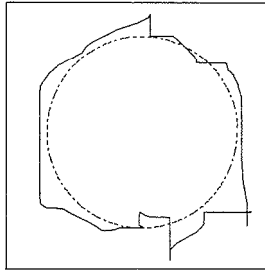


AN EXPERIMENTAL COMPARISON BETWEEN A LIGHT PEN AND A JOYSTICK WHEN USED FOR FREE-HAND DRAWING



LIGHT PEN



JOYSTICK

by

Niels Henrik Frey

DAIMI PB-100

August 1979

<p>Computer Science Department AARHUS UNIVERSITY Ny Munkegade - DK 8000 Aarhus C - DENMARK Telephone: 06 - 12 83 55</p>	
--	--

ABSTRACT.

This paper reports the results of an experimental study, made by Henrik Andersen and the author at the Department of Computer Science of Aarhus University as a part of their Masters' Thesis work.

A comparison is made between a light pen and a joystick, when used for free-hand drawing on a CRT-screen. The joystick proved superior for drawing horizontal and vertical lines, but hardly for anything else. The light pen proved to be an all-round instrument for free-hand drawing.

The subjects of the experiment were 52 undergraduate students at the Department, each of whom carried out two sessions of drawing a predefined set of images on the screen. After each session the students filled in a questionnaire on their own evaluation of the resulting images.

The paper contains a discussion of some of the methodological problems which arise in the planning and execution of investigations into man-computer interaction.

TABLE OF CONTENTS.

INTRODUCTION.	4
THE EXPERIMENT.	6
THE DRAWING INSTRUMENTS.	6
THE COURSE OF A SESSION.	8
THE MEASURED QUANTITIES.	10
THE USER GROUPS.	11
THE RESULTS OF THE EXPERIMENT.	11
User evaluations.	13
Drawing deviations.	14
Drawing speeds.	16
CONCLUSIONS ON THE RESULTS OF THE EXPERIMENT.	17
Characteristics of the drawing instruments.	17
The drawing deviation versus the drawing speed.	18
The template types and the drawing modes.	18
METHODOLOGICAL PROBLEMS.	20
THE MODEL OF THE EXPERIMENT.	20
Components of the experiment.	21
Analysis of the user.	22
Inter-connected quantities.	25
THE CREDIBILITY OF THE RESULTS.	28
The validity of the experimental results.	29
The accuracy of the experimental results.	30
Computational demands on the measures.	31
THE CHOICE OF USERS.	32
THE PRELIMINARY INVESTIGATION.	33
CONCLUSION.	34

1. INTRODUCTION.

It has often proved difficult to evaluate the performance of computer systems based on interactions between computers and humans. The problem seems to lie mainly in the trade-off between computer system expenses and user effectiveness. Some research has been done in this area, mostly on the problem of deciding when to use batch mode and when to use on-line mode in computer processing. Several investigations on the batch versus on-line topic were carried out during the period of 1966-1971. Most of these investigations were concerned with the effect of computer system response times on user performance.

Recently (since 1976) several papers on man-computer interaction has appeared. Some of these papers still treat the effect of computer system response time on the user performances, while many others deal with the effects of more general interactive system qualities on user problem solving abilities.

[Miller77] gives a survey of many of the investigations and experiments conducted up to 1977.

The results of these investigations and experiments have been characterized by several points of resemblance.

1. It has been difficult to generalize the results of the specific experiment to the whole problem area of man-machine interaction. The problem area is large and complex, and most experimenters have chosen to treat only restricted topics. Furthermore it has proved difficult to draw up mutually comparable criteria for the performance of the total systems.
2. No comprehensive methodology for planning and executing such investigations seems to exist [Sackman70].
3. The users of the system have always shown large individual differences in their personal performance. In many of the experiments it has been difficult to distinguish between measuring the overall performance of a man-machine system, and measuring the individual abilities of users.

This paper reports on an experiment which has much in common with the above mentioned investigations. The experiment deals with man-machine interaction, where both system characteristics and the individual characteristics of the involved users play an important role.

Evidently the problems which we have considered are of a restricted kind. Neither the planning nor the execution of the experiment have followed any predefined method, as no such thing seemed to exist at the start of the experimental work. But great attention has been paid to the individual differences of the involved persons, and we have spotted the largest methodological problems in planning and executing an experiment of this sort.

This paper therefore gives some contributions to the field of investigations of interactive computer systems.

2. THE EXPERIMENT.

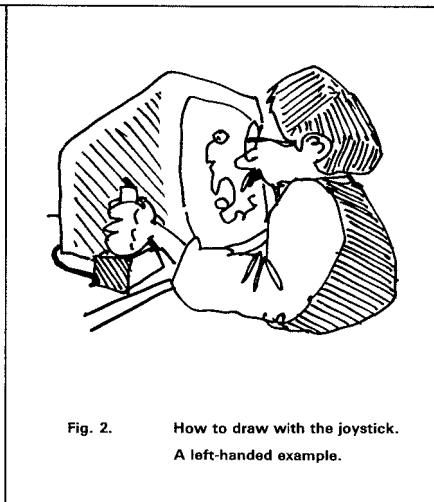
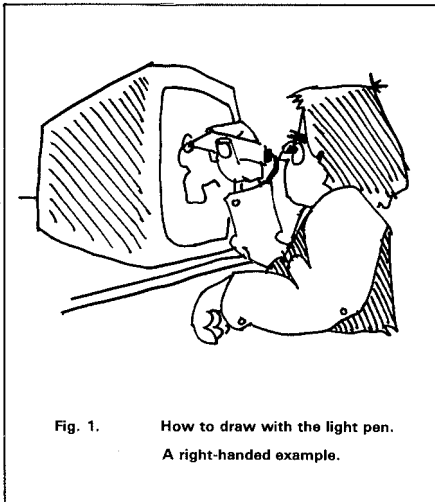
The experimental investigation of the two drawing instruments has arisen from a need to evaluate computer equipment using more than intuitive feelings. As mentioned before such evaluations are particularly difficult if the use of the equipment involves continuous interaction with human beings.

In this chapter we shall provide a short summary of the experiment and the results related to the drawing instruments.

2.1. THE DRAWING INSTRUMENTS.

The light pen used in the experiment, has been manufactured at the Department of Computer Science in Aarhus. It is a light pen of ordinary type with a switch, which when depressed makes the light pen sensitive to light emitted from the CRT of the graphics screen.

The joystick too has been manufactured locally at the Department. The joystick is in the neutral position, when the handle is held vertically. The handle is spring-loaded towards the neutral position by two spring-systems at right angles to each other. As a consequence, two directions at right angles to each other (normally perceived as "vertical" and "horizontal" on the screen) are very easy to draw with the joystick. One of the purposes of the experiment is to measure the consequences of this effect on free-hand drawing.



The graphics screen, on which the experiment was conducted, is a DEC GT-40 containing a small computer (a PDP-11/05) and a CRT-screen, and connected to a time-shared computer (a PDP-10).

The two drawing instruments are normally used in two basically different ways. If one imagines a "drawing point" on the graphics screen, the light pen will be used to point out directly on the screen the next position for the drawing point. This resembles the normal method of free-hand drawing if it is done continuously. See fig. 1.

The joystick, however, is used to indicate the directions in which the drawing point should move. By a smaller or larger deflection of the joystick handle from the neutral position, the user is able to indicate the speed with which the drawing point moves. This is a somewhat unorthodox method of free-hand drawing. See fig. 2.

In other words, with the light pen one can draw in "absolute coordinates", while one can draw in "relative coordinates" with the joystick.

The purpose of the experiment has been to evaluate the characteristics of the two drawing instruments when used for free-hand drawing. We chose free-hand drawing primarily because we found this method of man-machine interaction highly interesting. Secondly because free-hand drawing requires great control over the drawing instrument, and thus we expected many of the characteristics of the drawing instruments to show up in the experimental results.

From our own experiences with the instruments, we predicted that their drawing characteristics would be as shown in fig. 3. This prediction proved to be rather inaccurate.

The light pen	Poor performance at straight lines.	Good performance at rounded curves.
The joystick	Good performance at straight lines.	Poor performance at rounded curves.

Fig. 3. The preliminary judgement of the drawing instruments.

2.2. THE COURSE OF A SESSION.

In the experiment, each user went through two identical drawing sessions, with at least one day between them, to lessen the influence of the first session on the second session. Some users employed the same drawing instrument in both sessions, while some employed a light pen in one session and the joystick in the other.

In each drawing session two aspects of the characteristics of the drawing instruments were investigated: Their drawing properties on different templates and their drawing properties in different drawing modes. Each user made drawings based on three different types of templates, see fig. 4. Furthermore the users drew in two different modes, see figs. 5 and 6.

Type a:	Templates consisting exclusively of horizontal and vertical lines.
Type b:	Templates consisting of straight lines neither horizontal nor vertical.
Type c:	Templates consisting of rounded curve-segments.

Fig. 4. The three template types used in the drawing experiment.

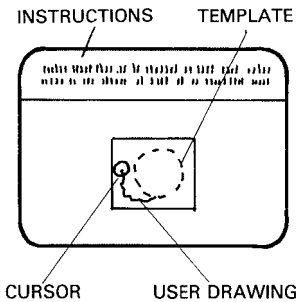


Fig. 5.

The picture on the screen in drawing mode 1.

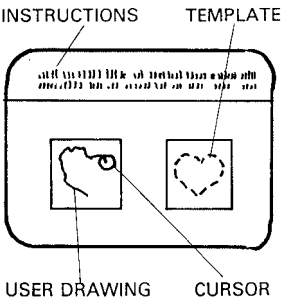
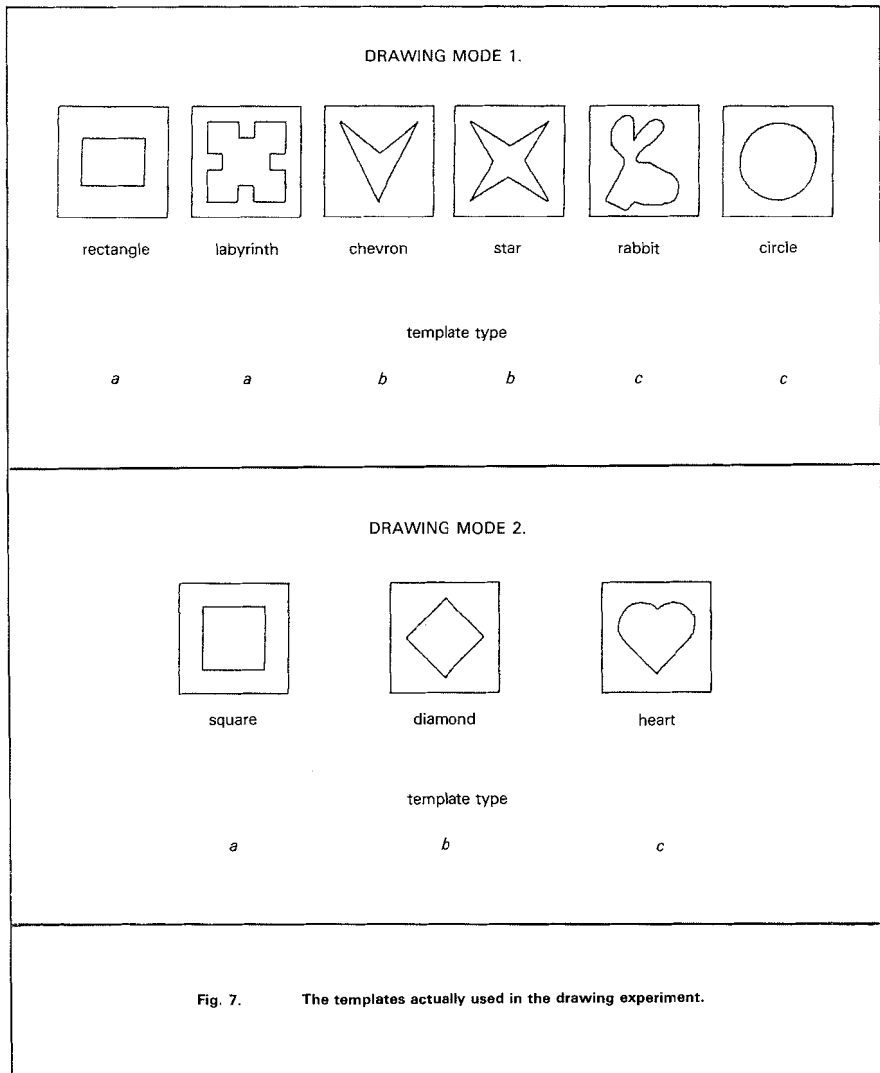


Fig. 6.

The picture on the screen in drawing mode 2.

In the first mode a user would ink in the template on the graphics screen. See fig. 5 for an example of the picture on the screen. By moving the cursor (and with that the drawing point) with the light pen or the joystick, he was able to make his drawing directly on top of the template. In this drawing mode the user got direct visual feed back on the resemblance between the template and his drawing.



In the second drawing mode the user would see the template sitting on one part of the screen, and make a copy on another part of the screen. See fig. 6 for an example of the picture on the screen. Using this drawing mode it was necessary for the user to maintain a mental picture of the size and proportions of the template. It was always possible for the user to refresh this mental picture by looking at the template.

Based on several criteria concerning the layout of templates (among others a high degree of symmetry), the templates in fig. 7 were used. The templates in drawing mode 2 had to be extremely simple to ensure that the user could hold a mental image of them. The rabbit-template is a bit special, as it does not fulfil the criterion of symmetry. Besides it is the only template which has not been "programmed". It was drawn by the experimenters using a light pen. The rabbit-template was included in the session, partly because we liked it, and partly because we wanted to see, whether the results from an untraditional template would differ much from the results from the other templates.

Each drawing session consisted of the user making his copies of the above mentioned nine templates.

2.3. THE MEASURED QUANTITIES.

In each drawing session three quantities were measured for each user drawing.

1. A quantity called the *user evaluation* of the drawing. After each drawing session the user would fill in a questionnaire concerning his satisfaction with his drawings. For each drawing he would have the choices:
VERY GOOD / GOOD / OK / NOT VERY GOOD / BAD.
2. A quantity called the *drawing deviation*, computed as the area between the template and the user drawing, divided by the line-length of the template.
3. A quantity called the *drawing speed*, computed as the line-length of the user drawing, divided by the time taken to make the drawing.

These three quantities were selected among several possible candidates, mainly because of their easily understandable information content, and because of their relative ease of computation.

2.4. THE USER GROUPS.

The people who participated in the drawing experiment as users, were second-year students, all following the same course in Computer Science at the Department. These students had received a comprehensive introduction to Computer Science, but they had not tried to work with the two drawing instruments, nor with graphics screens (the consequences of this inexperience will be discussed later in this paper). Participation in the experiment was voluntary and gave neither course credit nor any payment.

Following an introductory lecture on the experiment and its background, fifty-two students volunteered as users, and these students were randomly divided into four user groups. The members of each user group went through the two drawing sessions with the drawing instruments as shown in fig. 8.

	group 1	group 2	group 3	group 4
1st session	light pen	light pen	joystick	joystick
2nd session	light pen	joystick	joystick	light pen

Fig. 8. Table of the drawing instruments employed by members of the four user groups in the two drawing sessions.

2.5. RESULTS OF THE EXPERIMENT.

When the data was processed, the results from nine persons were discarded because the time-shared computer which controlled the course of each drawing session and collected data from the graphics screen, crashed during one of their drawing sessions. The results of one user were rather extreme, and they were also discarded, as they clearly ruined the information content of the results from the rest of the user group.

The following results have been calculated from user groups of the sizes shown in fig. 9. Because of great individual differences in the drawing abilities of the users, only results calculated from differences between two sessions can be deemed trustworthy. The reason of this is shown in section 3.1.2. The differences are computed as (*results from second session - result from first session*).

group 1	group 2	group 3	group 4
10	11	11	10

Fig. 9. The number of users in each user group from which results have been calculated.

We shall concentrate on two of the three measured quantities mentioned in section 2.4, namely the user evaluations and the drawing deviations. The results calculated from drawing speed do not show any obvious patterns, probably because the users concentrated on getting the drawings to resemble the templates as much as possible, not caring how long the drawing process was taking.

2.5.1. USER EVALUATIONS.

Because of problems in calculating the difference between two user evaluations (what is the difference between "VERY GOOD" and "OK" ?), we have chosen to classify the differences between the first and the second session in three groups: progress, no difference and regress. The differences of user evaluations between first and second session is shown for the four user groups in figs. 10 and 11.

In fig. 10 user evaluations are shown for user groups having used the same drawing instrument in both sessions. There is a slight tendency to evaluate the drawings from the last session highest.

In fig. 11 user evaluations are shown for user groups having employed different drawing instruments in the two sessions. There is a pronounced tendency to evaluate the joystick better at template type *a* drawings in both drawing modes, while the light pen is clearly evaluated as better at template type *c* drawings in both drawing modes. Apparently the evaluation of template type *b* drawings depends on the drawing mode.

Note that the two sets of evaluations are mutually consistent. It does not seem to make any difference, which drawing instrument has been used first.

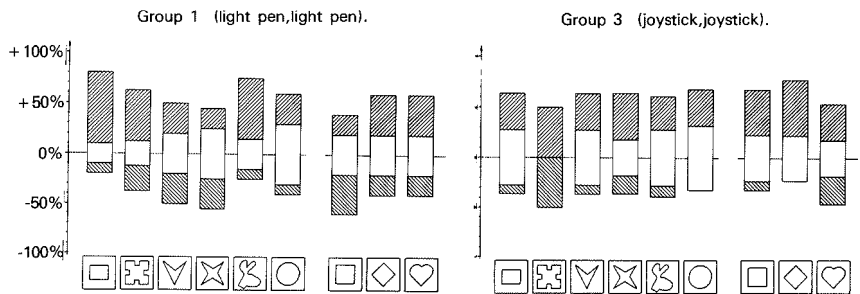





Fig. 10. Differences between user evaluations from 1st to 2nd session. The results are from user groups employing the same drawing instrument in both sessions. Differences are categorized in the following classes:

-  Progress from 1st to 2nd session.
-  No difference between 1st and 2nd session.
-  Regress from 1st to 2nd session.

For each template the percentage of progress, no difference and regress is shown.

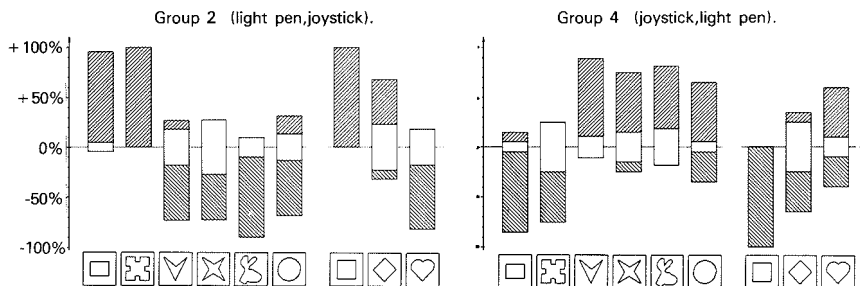





Fig. 11. Differences between user evaluations from 1st to 2nd session. The results are from user groups employing different drawing instrument in each session. Differences are categorized in the following classes:

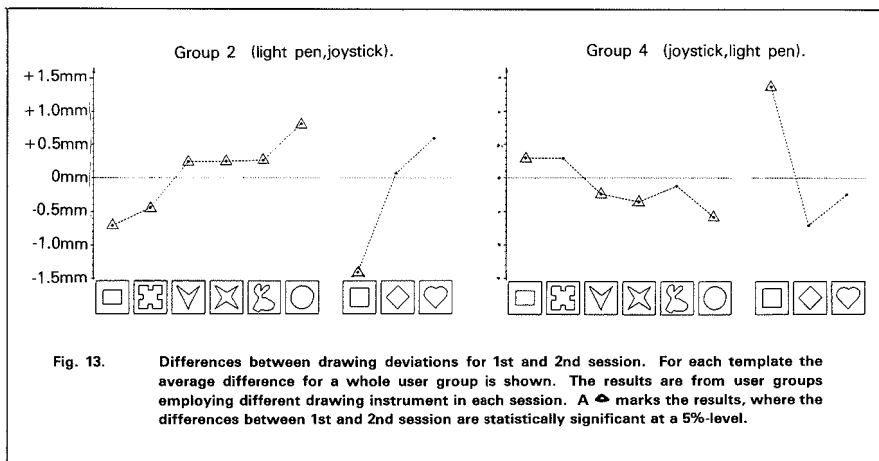
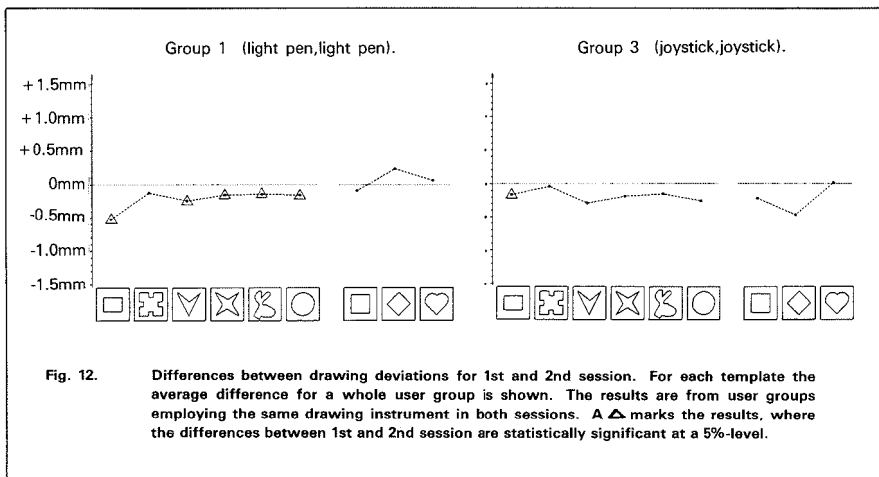
-  Progress from 1st to 2nd session.
-  No difference between 1st and 2nd session.
-  Regress from 1st to 2nd session.

For each template the percentage of progress, no difference and regress is shown.

2.5.2. DRAWING DEVIATIONS.

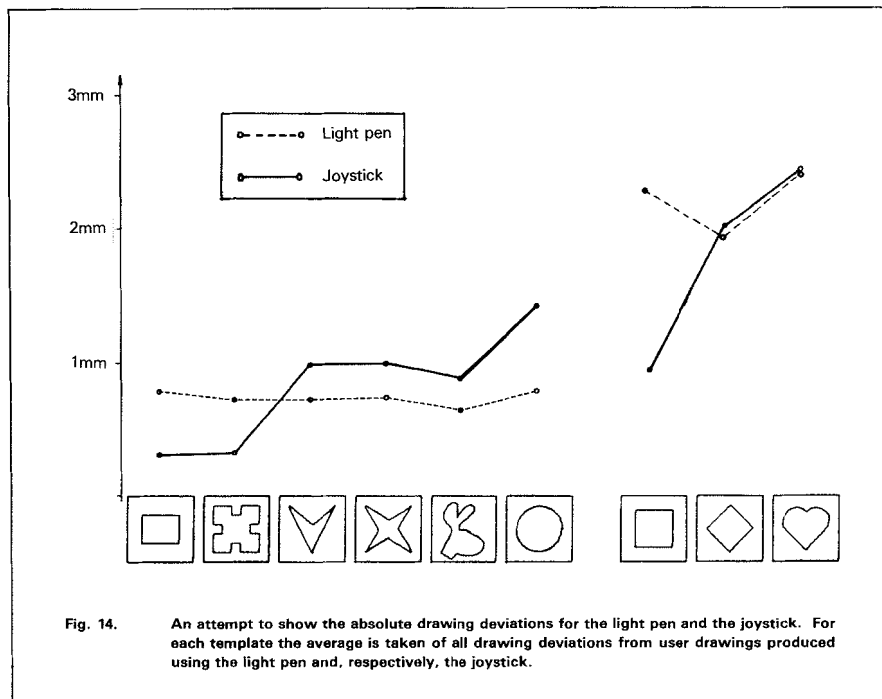
For each user group the personal differences between the drawing deviations of the first and the second session has been calculated and statistically analysed. In figs. 12 and 13 the averages of these differences is shown for each user group. As a small drawing deviation is considered good, the reader should beware that a progress from first to second session will result in a negative difference.

Results from the two user groups employing the same drawing instrument in both sessions are shown in fig. 12. There is a learning effect in both groups, most markedly with the light pen. As the results from drawing mode 2 mostly show a large variation, it is difficult to attach much meaning to the averages shown in figs. 12 and 13 for this drawing mode.



In fig. 13 the drawing deviation differences from user groups employing different drawing instruments in the two sessions are shown. The drawing deviation differences show much the same pattern as the user evaluations in chapter 2.5.1 do. The joystick shows better performance at template type *a* drawings, while the light pen seems superior at template type *b* and *c* drawings. Note also, that these differences show the same consistency as the user evaluations. It does not seem to matter which of the drawing instruments the user starts with.

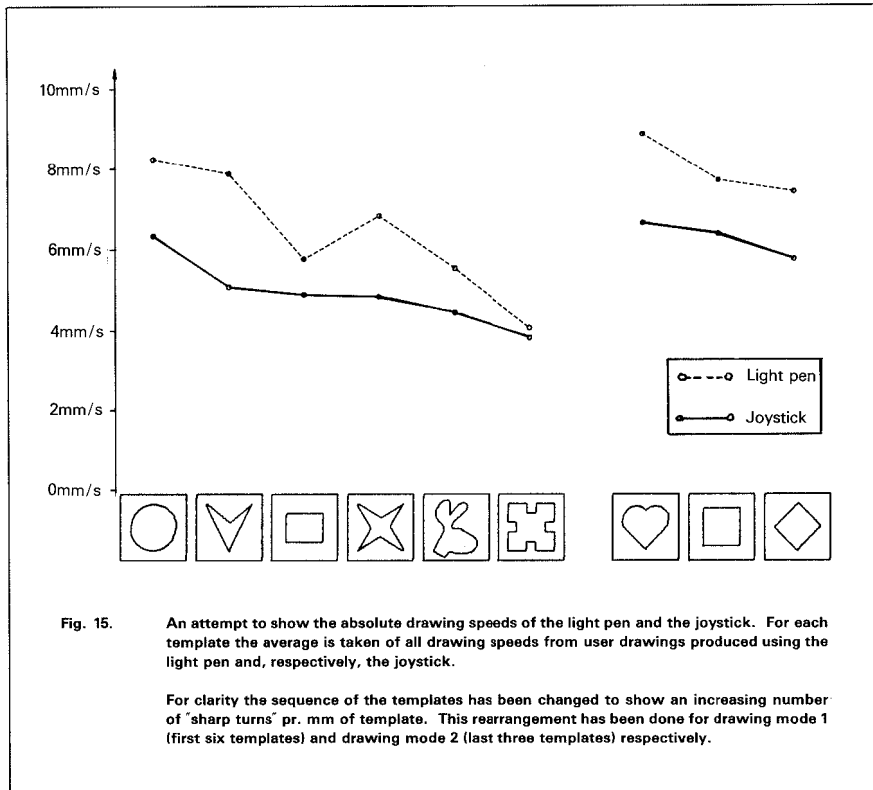
Even though it is nice to be able to tell the *differences* between the two drawing instruments, it would be somewhat nicer to obtain *absolute* estimates of their performance. In chapter 3.1.2 we shall discuss why it would be statistically infeasible to calculate such estimates. Nevertheless, as all drawing deviations from all drawings made with the light pen follow the same pattern, and as the same applies to the drawings made with the joystick, we have calculated two averages for each template. One average is made for all drawings made by the light pen. One average is made for all drawings made by the joystick. These results are shown i fig. 14.



2.5.3. DRAWING SPEEDS.

The drawing speeds measured in the experiment do not show a clear dependency upon drawing instrument, drawing mode and template as the drawing deviations do. As mentioned before, this might be a consequence of the users concentrating only on the goodness of the drawing, and not worrying about how long it took.

The differences between the drawing speeds from the first and second sessions do not show a coherent picture. In addition it is, for reasons which we shall not discuss here, not possible to apply any reasonably straightforward statistical analysis to the drawing speeds. This all shows that too little care has been taken in the planning of the experiment regarding the drawing speeds.



As the drawing speeds from all drawings made with the light pen approximately follow the same pattern, and as this also applies to drawings made with joystick, we have calculated averages for each template for all light-pen-made drawings, respectively all joystick-made drawings. These averages are shown in fig. 15. Note that no statistical significance can be attached to these figures. Note also that the order of the templates has been changed to give a clearer picture.

Three tendencies seem to be evident. Firstly that the light pen is faster than the joystick. Secondly that the difference in drawing speeds between the two drawing instruments decreases as the complexity of the template increases. Thirdly that drawing mode 2 is faster than drawing mode 1 (but also more inaccurate, as fig. 14 shows).

2.6. CONCLUSIONS ON THE RESULTS OF THE EXPERIMENT.

There are many conclusions that can be made on an experiment like this. The conclusions pertaining to the characteristics of the drawing instruments are made in sections 2.6.1 - 2.6.3. The conclusions about the way the experiment was planned and executed, are gathered in chapter 3.

2.6.1. CHARACTERISTICS OF THE DRAWING INSTRUMENTS.

The light pen seems to be an all-round free-hand drawing instrument with approximately the same drawing deviation for all types of figures (see fig. 14). The joystick seems to be very good at drawing horizontal and vertical lines, but hardly usable for anything else. The drawing performance of the joystick seems to get worse, the more rounded curves there are in the figures. Unfortunately these estimates of the drawing instruments are based on results shown in fig. 14, and those results are not renderable to statistical analysis.

Note that this evaluation of the performance of the light pen and joystick does not correspond to our original estimates. For comparison, see figs. 3 and 16. Note also, that the averages in fig. 14 give no specific indication of the "real" characteristics of the drawing instruments. We lack a norm to indicate whether a drawing deviation of, say 2 mm, is "good" or "bad". This problem (called the validity problem) is treated in more detail in section 3.2.1.

The light pen is faster to use than the joystick. This, together with the difference in deviations indicates that the joystick should only be used in free-hand drawing when this consist exclusively of horizontal and vertical lines. As this will almost

never happen in free-hand drawing, we conclude that the light pen examined is superior to the joystick examined at free-hand drawing. However, the superior performance of the joystick for horizontal and vertical lines suggests that it will be useful in drawing various sorts of diagrammatic drawings.

Light pen	"Good" at all sorts of drawings	
Joystick	"Very good" at vertical and horizontal lines.	"Bad" at everything else.

Fig. 16. The final judgement of the drawing instruments.

2.6.2. THE DRAWING DEVIATION VERSUS THE DRAWING SPEED.

Not surprisingly a connection shows between the drawing deviations and the drawing speeds, such as: *"A large drawing speed is connected with a large drawing deviation"*. And: *"A small drawing speed is connected with a small drawing deviation"*.

Unfortunately we were not aware of this (in retrospect extremely obvious) connection, and did not plan the experiment to investigate this. As a consequence, the results in figs. 12-15 are not as useful as they might appear, as the drawing deviations and the drawing speeds cannot be viewed separately, but must be perceived as two sides of the same entity. It has not been possible after the execution of the experiment to ascertain the relationship between the drawing deviations and the drawing speed.

2.6.3. THE TEMPLATE TYPES AND THE DRAWING MODES.

The nine templates used in the experiment seem well chosen, as they have been able to provoke different performances from the light pen and the joystick. The most astonishing fact in this connection is probably that the rabbit-template (originally drawn with the light pen) resulted neither in extremely good light pen results nor in extremely bad joystick results. As a matter of fact, the joystick performed better on the rabbit than should be expected from the results of the other templates (see figs. 11,13,14 and 15). We do not know how to explain this phenomenon.

A comparison between the two drawing modes is difficult because of the large variation in the individual results in drawing mode 2. Nonetheless it is possible to conclude, that the user draws faster and more inaccurate in drawing mode 2. Also the conclusions pertaining to the influence of the templates on the user drawings seem consistent for both drawing modes, but note that almost all the conclusions have actually been made from results from drawing mode 1.

3. METHODOLOGICAL PROBLEMS.

In this chapter the main problems of planning and executing an experiment like ours are treated. We shall concentrate on four issues:

1. The significance of making a model of the experiment and using this model in the analysis of the experimental situation.
2. The importance of controlling the translation process from what we want to investigate to what we actually measure.
3. The consequences of choosing different sets of users for the experiment.
4. The importance of making a preliminary investigation.

Examples from the experiment are given for each of the topics.

These issues are not the only ones involved in the making of an experiment, but they are the ones we deem most important.

3.1. THE MODEL OF THE EXPERIMENT.

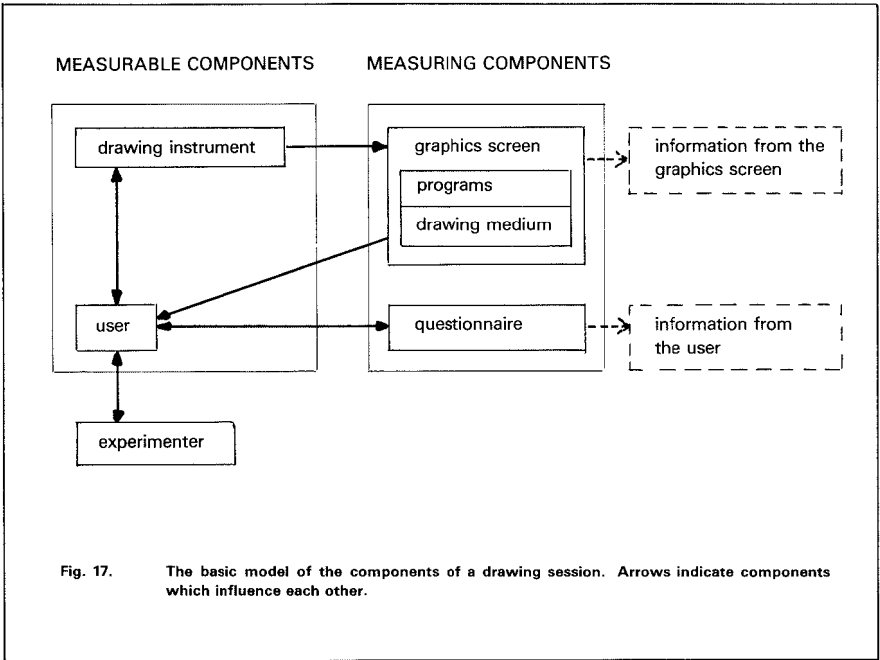
An important tool in the analysis of the experimental situation is a model, which consists of the components of this situation and their interaction. A reasonable model will help the experimenter to plan and execute his experiment, as it will help him to distinguish between measuring relevant and irrelevant characteristics of the components of the experiment. As regards the relevant characteristics, he must decide whether or not they can be measured, and, if so, how this can be done. As regards the irrelevant characteristics he must analyze whether they exert a direct or indirect influence on the quantities he wants to measure, and he must try to arrange the experiment so as to minimize this influence.

3.1.1. COMPONENTS OF THE EXPERIMENT.

Fig. 17 shows the five most important components of a drawing session. The arrows between components indicate some sort of influence between the respective components. An analysis must now be made to show, how these influences will affect the quantities that the experimenter wants to investigate. The components have been divided in three groups: the *measurable components*, the *measuring components* and the *experimenter*.

The measurable components are those which we wish to investigate. It is often difficult or dangerous to alter their characteristics, as this may influence the results of the experiment in unexpected ways. But it normally is possible to pay regard to them in the planning of the experiment.

The measuring components are used to measure characteristics of the measurable components. These measuring components must be constructed so as to measure relevant characteristics and suppress irrelevant characteristics of the measurable components.



In fig. 17 the experimenter has been placed outside both the group of measurable components and the group of measuring components. He ought to be placed as one of the measuring components, but, because of an early decision in the planning of the experiment, we chose to let "him" play as little a role in the drawing sessions as possible. We were not able to measure the impact on the experimental result from the interaction between user and experimenter if the experimenter acted as an observer. (In [Hansen78] observation of the users proved to give large influences on the measured results.) As it was necessary for the experimenter to attend the drawing sessions in case of machine failures, we chose the status of the experimenter to be a person present, but not observing. In other words, the results of an interaction between experimenter and user was deemed irrelevant, and the non-observing status of the experimenter was the (rather crude) method used to minimize this interaction.

Fig. 17 helps to understand the factors involved in the experiment, but it has three shortcomings:

1. The internal structure of the components cannot be shown unless we lose the general view of the experimental situation.
2. It does not show the kinds of influences between the components.
3. The diagram give a somewhat static description of the drawing session - important dynamic features in such a session might not show.

So it is necessary to conduct several detailed analyses in connection with fig. 17, where each analysis concentrates on important aspects of the experiment.

3.1.2. ANALYSIS OF THE USER.

As an example of a detailed analysis, we shall show how a model of the characteristics of a user will influence the layout of the drawing experiment.

Fig. 18 lists the user characteristics which will presumably affect the data measured from the process of making a user drawing and a user evaluation. We restrict ourselves to describing the analysis of the characteristics influencing the user drawing. These are listed in fig. 19.

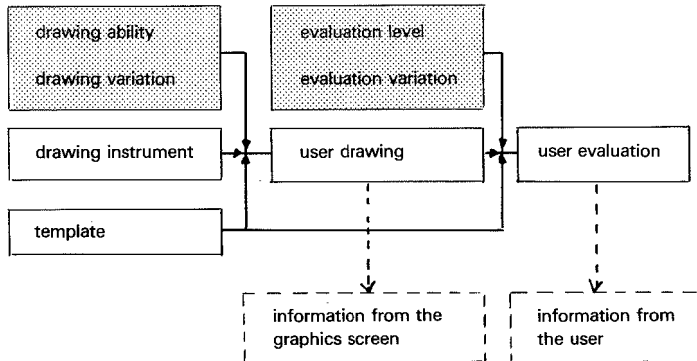


Fig. 18. Some characteristics of the user which influence the results of the experiment.

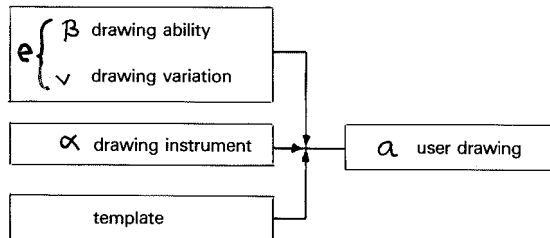


Fig. 19. The characteristics affecting the user drawing for a given template.

Three factors have been sorted out:

1. The *drawing ability* of the user. It gives a measure of the user's general capability at free-hand drawing.
2. The *drawing variation* of the user. It gives a measure of the difference between a user's actual performance and his drawing ability. As users are never capable of reproducing the same drawing precisely again and again, it is necessary to include such a factor in the analysis.
3. The *drawing instrument* and the *template*. This term shows the combined effect of the drawing instrument and the template used to make a user drawing. They have been considered one entity here to simplify the example of analysis.

Note that the drawing abilities are dynamic variables, which are liable to change value in the course of the experiment. The drawing variation are stochastic variables with an assumed mean value of zero. In fig. 20 the strict denotations of the variables in the case of one, respectively two, sessions in an experiment are shown.

With these denotations we can now give a more detailed model of how the user drawing is influenced by the user, the drawing instrument, and the template. In fig. 20 a simple additive context of the factors is shown for one session, respectively two sessions, per user in the experiment.

In fig. 21 two different ways of showing differences in the performance of drawing instruments are shown. The problem is well known in statistics as the "paired" (two sessions/one user group) versus the "unpaired" (one session/two user groups) case. From the statistical analysis the following can be concluded:

1. The investigation of the drawing instruments must be made with reasonably large, homogeneous groups to ensure that the variations on the measurements from the user drawing abilities and the user drawing variations do not make the testors ineffective in a statistical sense.
2. The investigation of the drawing instruments must include two drawing sessions for each user. The testors in fig. 21 show that large variations in the users' drawing abilities will make it almost impossible to get statistically significant answers from tests made on the one-session results. As we can easily assume that the users

have varying drawing abilities, it is a fundamental demand that the experiment contain two drawing sessions for each user. This gives as a result that the statistical analysis can only be expected to yield results when applied to *differences* between performances made in two sessions, as mentioned in section 2.5.2.

3. It is essential to let user groups draw with the same drawing instrument in both sessions to get a measure of the stability of the users' drawing abilities (in order to make the assumption of no learning effect in fig.20).

Some of the results of the demands, which this analysis has made upon the layout of the experiment, can be seen in fig. 8. This analysis has also influenced the choice of users, see section 3.4.

In fig. 18 we showed two characteristics of the user which deal with his evaluation of drawings when he filled out the questionnaire. An analysis similar to the previous one has been conducted upon these characteristics. The results have influenced the questionnaire and how the users should fill it in.

3.1.3. INTER-CONNECTED QUANTITIES.

Lastly we shall point out an important aspect of the making of a model which we have somewhat neglected when making the drawing experiment. It is extremely important to analyse how quantities measured in the experiment affect each other. In the analysis described on the previous pages importance has only been attached to the direct or indirect influence of other factors on the measured quantities.

In the drawing experiment, a drawing deviation and a drawing speed were measured for each user drawing. In the presentation of the results from the experiment, sections 2.5.1 - 2.5.3, these quantities are viewed independently of each other. As mentioned in section 2.6.2 there is a strong correlation between these quantities, and this correlation tends to lessen the applicability of the results presented for each of the quantities.

The drawing sessions ought to have contained possibilities for comparison of drawing deviations and drawing speeds. The users could have tried to draw some templates both quickly and slowly (but this would also have enlarged the experiment considerably).

DENOTATIONS.

One Session / Two User Groups.	Two Sessions / One User Group.
a_{jk} A measure on a drawing made in user group j by user k .	a_{tk} A measure on a drawing made in session t by user k .
α_j Influence from the drawing instrument determined by the user group number j .	α_t Influence from drawing instrument determined by session number t .
e_{jk} Stochastic variations on the measure, caused by user drawing ability and user drawing variation. With only one session, it is impossible to distinguish between these two user quantities.	β_{tk} Influence from user drawing ability of user k in session t .
	v_{tk} Stochastic variations on the measure, caused by user drawing variation from user k in session t .

BASIC ASSUMPTIONS.

One Session / Two User Groups.	Two Sessions / One User Group.
$a_{jk} = \alpha_j + e_{jk}$ where $j \in J = \{1, 2\}$ and $k \in K_j = \{1, 2, \dots, n_j\}$. $e_{jk} \sim N(0, \sigma_j^2)$.	$a_{tk} = \alpha_t + \beta_{tk} + v_{tk}$ where $i \in I = \{1, 2\}$ and $k \in K = \{1, 2, \dots, n\}$. $v_{tk} \sim N(0, \sigma^2)$. <p>If we further assume that a shift from one drawing instrument to another produce no learning effect on the user's drawing ability, we may have $\beta_{1k} = \beta_{2k} \quad \forall k \in K$.</p> <p>Then $a_{tk} = \alpha_t + \beta_k + v_{tk}$ and $v_{tk} \sim N(0, \sigma^2)$.</p>

Fig. 20. Notations and basic assumptions of the detailed analysis of the user's influence on his drawing.

TESTING DRAWING INSTRUMENT INFLUENCES.

One Session / Two User Groups.

With the assumptions shown in fig. 20, it is easy to test the hypothesis $\alpha_1 = \alpha_2$.

A student-distributed testor t is:

$$t = \frac{\overline{a_1} - \overline{a_2}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

where $\overline{a_j} = \frac{1}{n_j} \sum_{k=1}^{n_j} a_{jk}$, $\forall j \in J$.

$$s^2 = \frac{1}{2}(s_1^2 + s_2^2) \text{ and } s_j^2 = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (a_{jk} - \overline{a_j})^2.$$

Two Sessions / One User Group.

With the assumptions shown in fig. 20, it is easy to test the hypothesis $\alpha_1 = \alpha_2$ (which is equivalent to testing $\alpha_1 + \beta_k = \alpha_2 + \beta_k \forall k \in K$).

A student-distributed testor t_d is:

$$t_d = \frac{\overline{d}}{s_d \sqrt{\frac{1}{n}}} \sim t(n-1)$$

where $d_k = a_{1k} - a_{2k} \forall k \in K$

$$\text{and } s_d^2 = \frac{1}{n-1} \sum_{k=1}^n (d_k - \overline{d})^2$$

A COMPARISON BETWEEN THE TWO TESTORS.

As both testors have the same numerator, the comparison will only concern the denominators, in effect only the s 's in these.

One Session / Two User Groups.

$$s_j^2 = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (a_{jk} - \overline{a_j})^2 = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (\alpha_j + e_{jk} - \alpha_j - \overline{e_j})^2 = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (e_{jk} - \overline{e_j})^2.$$

Two Sessions / One User Group.

$$s_d^2 = \frac{1}{n-1} \sum_{k=1}^n (d_k - \overline{d})^2 = \frac{1}{n-1} \sum_{k=1}^n (\alpha_1 + \beta_k + v_{1k} - \alpha_2 - \beta_k - v_{2k} - \alpha_1 - \overline{\beta} - \overline{v_1} + \alpha_2 + \overline{\beta} + \overline{v_2})^2 \\ = \frac{1}{n-1} \sum_{k=1}^n ((v_{1k} - \overline{v_1}) + (v_{2k} - \overline{v_2}))^2.$$

Note that the s_j include variations from both the user drawing abilities and the user drawing variations, while the s_d include variations from the user drawing variations only. So if large individual differences in user drawing abilities are to be expected, the s_d will probably be smaller than the s_j . The testor with the s_d will therefore be more apt to show differences between drawing instrument performances, as a large tester will overthrow the hypothesis of no difference (i.e. $\alpha_1 = \alpha_2$).

Fig. 21. Statistical testors to investigate differences between measurements, respectively for comparison between two different groups and for comparison between two sessions done by the same group.

3.2. THE CREDIBILITY OF THE RESULTS.

Another aspect of the extreme importance in the planning of an experiment, is the relation between what one wants to measure, and what is actually measured. In fig. 22 is shown an ordering of measures in three levels.

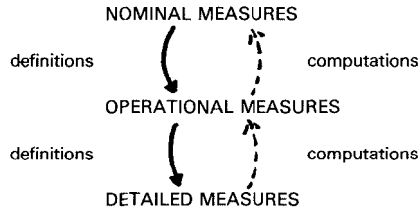


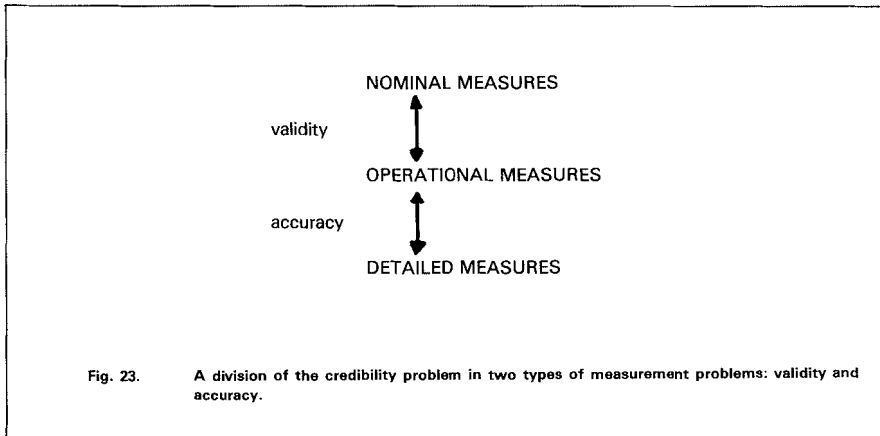
Fig. 22. A three-level schema of the measures used in an investigation.

The *nominal measures* are those which we really want. In the case of the drawing experiment, we wanted a measure of **applicability at free-hand drawing** for the drawing instruments. As it is seldom possible to apply such measures directly (as no corresponding measuring instruments exists), it is necessary to translate the nominal measures into operational measures. If this was not necessary, we would hardly feel any need for an experiment.

The *operational measures* are characterized by being quantifiable in some fashion, and it is therefore easier to construct measuring instruments for them. The operational measures of the drawing experiment are the three quantities mentioned in section 2.3: a one-dimensional **user evaluation**, a one-dimensional **drawing deviation** and a one-dimensional **drawing speed**. The operational measures are in many ways the most important measures, as the results of the whole experiment are statistical interpretations of these operational measurements.

Though the operational measures are quantifiable, it is usually a non-trivial task to settle on which quantities to measure on a detailed scale - the *detailed measures*. So it is important to check the correspondence between the operational measures and the detailed measures. In case of the drawing experiment the following information was collected: The users' evaluations of the drawings, chosen from **five standard ratings** (see section 2.3), all the **lines** in the user drawings, and the **time** taken to make the drawings.

There are four translation problems, as depicted in fig. 22, where the two translations which define measures are the most important. In fig. 23 the two measure translations, which together establish the credibility of the results from the experiment, are shown.



3.2.1. THE VALIDITY OF THE EXPERIMENTAL RESULTS.

In establishing the validity of the measuring instrument, we try to determine whether or not we actually measure what we want to measure. We want to see if there are any systematic errors of measurement in the results.

As the subject of validity problems has been thoroughly treated in many introductory sociology books, we shall only describe, what was done in the drawing experiment. The user evaluations and the drawing deviations were compared, as we wanted them to be two ways of measuring the applicability of the drawing instruments for free-hand drawing. A nice agreement between the two sets of measurements would prove nothing, as the measurements might be systematically wrong in the same way. But a disagreement between the two sets of measurements would show that at least one of the operational measures was wrong.

A correlation analysis of the two sets of measurements showed a nice agreement. As the two operational measures came from two quite different sources (human beings and machines), the agreement was nonetheless interpreted as indicating a rather high validity for the experimental results. This also indicates, that the problems rising from the interconnection between the drawing deviations and the drawing speeds (see section 2.6.2) are not as serious as we might think.

It is important to note, however, that an agreement between two sets of measurements does not indicate that the corresponding measures are mutually compatible. In the case of the drawing experiment it would be an error to state that, for instance, the drawing deviations can be used on their own instead of the user evaluations, even if this experiment has shown them to be highly correlated. It is very likely that the shape of the user drawings has also had some influence on the user evaluations.

It is now clear that the need to investigate the validity of the results made a demand upon the planning of the test, as we found it necessary to measure the applicability of the drawing instruments in more than one way, resulting in both human and mechanical measures.

3.2.2. THE ACCURACY OF THE EXPERIMENTAL RESULTS.

In establishing the accuracy of the measuring instrument, one tries to determine whether or not it measures consistently, free from unsystematic errors. In other words, the measuring instrument is accurate if there is a large agreement between two or more measurements of precisely the same thing.

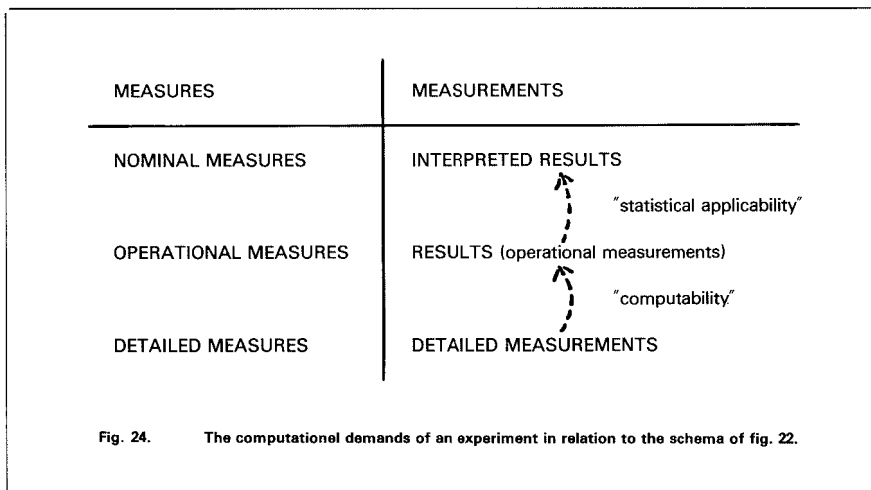
As with validity problems, the accuracy problems are well-known in sociology, and again we shall only present what was done in the drawing experiment to establish the accuracy of the measurements.

With human beings it is impossible to measure precisely the same thing twice, as he or she will never react in precisely the same way both times. So several measurements of the same person can only be done in a limited sense. If the four user groups are chosen randomly from a large, homogeneous pool of users, it can be shown that user groups with the same history in the experiment are mutually comparable. As these criteria are met in the drawing experiment, measurements from the first drawing session of user group 1 and 2, respectively user group 3 and 4, were compared (see fig. 8). A correlation analysis showed a very beautiful agreement between the drawing deviations and the drawing speeds, while the user evaluation measurements showed an acceptable agreement only. This indicates that the questionnaire should have been constructed with more care.

The need to investigate the accuracy of the measures in the experiment made demands on the planning of the experiments, namely that the set of users were homogeneous, and that the division of users into user groups was done randomly.

3.2.3. COMPUTATIONAL DEMANDS ON THE MEASURES.

When translating the nominal measure to operational measures to detailed measures, it is most important to establish the validity and the accuracy of the resulting measuring instrument. But the experimenter must also recognize problems related to the computational translation process back from the *detailed measurements* to the *results* to the statistical *interpretation of the results*, see fig. 24. Two small examples to enlighten this.



In the drawing experiment the drawing deviation was defined as the area between the template and the user drawing, divided by the line-length of the template. To avoid ambiguities and to avoid computational complications, a demand was made that the templates should not be self-intersecting, even though this excluded a large class of possible templates. The demand of the "computability" of the operational measurements thus has an effect on the experiment layout.

In section 3.1.2, some demands on the experimental design originated in an analysis of the characteristics of the users. Here statistical considerations had a large influence on the layout of the experiment. It was no coincidence that the model in fig. 20 was renderable to several kinds of simple statistical analysis (such as linear regression analysis and analysis of variance).

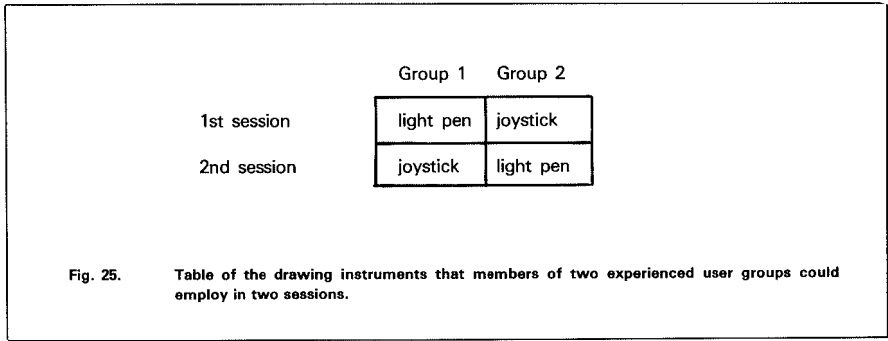
It is of course dangerous to make models applicable only to some statistical tools - as we learned later, when we found out that our statistical tools did not allow any comparison between drawing deviations and drawing speeds. But it is better than having results which are shown to be both valid and accurate but totally unfit for statistical analysis, either because no statistical tools are applicable, or because all of the results of the statistical analysis are insignificant.

3.3. THE CHOICE OF USERS.

The credibility of the results of an experiment like this is strongly connected to the choice of users. If the set of users is non-representative, little faith can be put in the experimental results. It is also important to clarify, which characteristics of the users can be expected to influence these experimental results, and to incorporate the characteristics in the model. In section 3.1.2, differences in the users' drawing abilities were used in the model.

In the drawing experiment it would have been possible to employ two sets of users. The first set was small and inhomogeneous, consisting of graduate students with some experience in working with graphics screens and the light pen and joystick. The other set was large and fairly homogeneous, consisting of second-year undergraduate students with no experience in the use of the drawing instruments, but with almost two years of experience in the use of general purpose computers.

The experienced users would probably show no learning effect, and so it would suffice to have two user groups, as shown in fig. 25. But the small size and the inhomogeneous character of the group would probably make it unrenderable to statistical analysis. As previously mentioned in section 3.3.2, the investigation of the accuracy of the measuring instrument would also be difficult.



The inexperienced users would, on the other hand, probably show significant learning effects. So it would be necessary to make four user groups as shown in fig. 8, where two of the user groups could be used as control groups to measure those effects. This also made the statistical analysis and the investigation of the accuracy of the measuring instrument easier. Another problem with this set of users is their inexperience seen in a more general view.

In the end we chose the large, homogeneous and inexperienced user group, hoping that the learning effects would not be serious, and that the tendencies derived from the results would also apply to more experienced users.

3.4. THE PRELIMINARY INVESTIGATION.

The applicability of the results from an experiment are strongly dependent upon the assumptions on which the experiment has been built. In other words, it is of great importance to be able to test the model of the experiment before it is too late.

So it is necessary to arrange a preliminary investigation. This investigation will normally disclose some technical imperfections of the experimental layout, but its primary function is still to test the applicability of the model of the experiment. All details in the preliminary investigation must therefore be as close to the final experiment as possible, with one exception: the user groups cannot be as large as in the real experiment. But the users in the preliminary investigation must be taken from the same population the "real" user groups are taken from. There is a basic flaw in every preliminary investigation, as every error in the experimental layout that only shows on large amounts of measurements, will be hard to locate. For instance the statistical distribution of measurements can be difficult to establish on few measurements.

As an example of the importance of the preliminary investigation, we use the model of the users' drawing characteristics in section 3.1.2.

The basic assumptions were:

$$a_{ik} = \alpha_i + \beta_k + v_{ik} \quad \text{and} \quad v_{ik} \sim N(0, \sigma^2).$$

If the preliminary investigation instead shows a multiplicative context like:

$$a_{ik} = \alpha_i \cdot \beta_k \cdot v_{ik} \quad \text{and} \quad \ln(v_{ik}) \sim N(0, \sigma^2).$$

Results based on differences between the first and second session will be useless, as the influences of the *user drawing abilities* (β_k) are not removed. Another layout of the experiment or other ways of measuring must be used. There is also the possibility that arithmetical transformations like logarithms on the data will result in an additive context again. But such transformations may alter the statistical properties of the measurements in nasty ways (this actually happened to the *user drawing speeds* in our experiment).

3.5. CONCLUSION.

Sections 3.1 - 3.4 do not constitute a complete methodology for experiments involving computers and human users. Only the most important topics have been treated.

We wish to point out that we have chosen to consider only one category of experiments investigating man-machine interaction. For simplicity we have only considered experiments experimenters plan and execute, and where users know very little about the experiment, so the users' knowledge will not influence the results.

Many of the problems of measuring user characteristics are still not clarified. Most of our experimental work has concentrated on "technical" problems in connection with the users (that is, how to avoid distorting already measured data). Too little emphasis has been put on why we measured certain quantities. As mentioned in section 3.1 there is a distinct possibility that the experimenter via his model "distorts" the measuring process. This can be done inadvertently from ignorance of experimental methodology or of the users' working situation, or it can be done on purpose to produce specific effects. So users co-operating with the experimenter during the whole experiment give large opportunities for checking the "applicabilities" of the experimental results.

Lastly it is important to note how the development of new uses for computers influences the computer science field. Not much of this paper can be classified as traditional computer science, but as the growing importance of close man-machine-interaction begins to show in many ways, several sciences previously deemed to have little to do with computer science, have become relevant to computer scientists. Many aspects of sociology, psychology (and statistics) have proved important in the drawing experiment, and will undoubtedly prove applicable in the analysis of the use of computerized equipment in the future.

REFERENCES.

- [Andersen79] H.C. Andersen and Niels Henrik Frey: **Test på et grafisk system.** (An experiment conducted on a graphics system). Master's Thesis. Department of Computer Science, Institute of Mathematics, University of Aarhus, Denmark, December 1978.
- [Boehm71] B.W. Boehm, M.J. Seven and R.A. Watson: **Interactive problem-solving - An experimental study of lockout effects.** Spring Joint Computer Conference 1971.
- [Carbonell69] Jaime R. Carbonell: **On man-computer interaction: A model and some related issues.** IEEE Transactions on Systems Science and Cybernetics, Vol. SSC-5, no. 1, January 1969.
- [Corley76] Melvin R. Corley and John J. Allan III: **Pragmatic information processing aspects of graphically accessed computer-aided design.** IEEE Transaction on Systems, Man and Cybernetics. Vol. SMC-6, No. 6, June 1976.
- [Cotton77] Ira W. Cotton: **Cost-benefit analysis of interactive systems.** Computer Networks 1. 1977.
- [Gannon77] J.D. Gannon: **An experimental evaluation of data type conventions.** Comm. of the ACM. Vol. 20, No. 8, August 1977.
- [Goodman78] T. Goodman and R. Spence: **The effect of system response time on interactive computer aided problem solving.** Siggraph-ACM, Vol. 12, No. 3, August 1978.
- [Grossberg76] Mitchell Grossberg, Raymond A. Wiesen and Douwe B. Yntema: **An experiment on problem solving with delayed computer responses.** IEEE Transactions on Systems, Man and Cybernetics, March 1976.
- [Hansen71] Wilfred J. Hansen: **User engineering principles for interactive systems.** Fall Joint Computer Conference 1971.
- [Hansen78] Wilfred J. Hansen, Richard Doring and Lawrence R. Whitlock: **Why an examination was slower on-line than on paper.** International Journal on Man-Machine Studies, No. 10. 1978, Academic Press Inc (London) Limited.

- [Kozar78] Kenneth A. Kozar and Gary W. Dickson: **An experimental study of the effects of data display media on decision effectiveness.** International Journal on Man-Machine Studies. No. 10, 1978, Academic Press Inc (London) Limited.
- [Miller68] Robert B. Miller: **Response time in man-computer conversational transactions.** Fall Joint Computer Conference 1968.
- [Miller77] Lance A. Miller and John C. Thomas Jr: **Behavioral issues in the use of interactive systems.** International Journal of Man-Machine Studies, No. 9, 1977.
- [Sackman67A] H. Sackman: **Time-sharing versus batch processing. The experimental evidence.** System Development Corporation, Santa Monica, California, 1967.
- [Sackman67B] H. Sackman: **Computers, system science and evolving society.** System Development Corporation, Santa Monica, California. Wiley and Sons, New York 1967.
- [Sackman70] H. Sackman: **Man-computer problem solving.** Auerbach Publishers Inc. 1970.
- [Schatzoff67] M. Schatzoff, T. Tsao and R. Wiig: **An experimental comparison of time sharing and batch processing.** Comm. of the ACM. May 1967.