Introduction to quantitative measurement

Psychophysics *vs.* Psychometrics

Christophe Lalanne
ch.lalanne@gmail.com

November, 2009

Summary

❝ Here we shall provide a general overview of measurement in quantitative psychology, with a special emphasis on psychomhysics and psychometrics theories. ❞

Outline

1. Psychophysics
2. Psychometrics
3. Test development and item analysis

Various functions used throughout this chapter were collated in the package **Psychomisc**.

Psychophysics *vs.* Psychometrics

These two broad fields in quantitative psychology have mereley evolved together until the mid 1970s.

The object of measurement

- Psychophysics focuses on
- Psychometrics centers on

Psychophysics: Some 'landmarks'

Most significant contributions: *Elemente der Psychophysik* (1860), *Handbook of Experimental Psychology* (1950), *The Measurement of Values* (1959/1974), *Signal Detection Theory and Psychophysics* (1966). . .

G. T. Fechner (1801–1887) W. Wundt (1832–1920) G. von Békésy (1899–1972) R. Likert (1903–1981)

E. H. Weber (1795–1878) S. S. Stevens (1906–1973) L. L. Thurstone (1887–1955)

Complete bibliographies may be found at `http://psychclassics.yorku.ca/`.

Weber's law

The Weber-Fechner law was first proposed by Weber (1795–1878) and later developed by Fechner (1801–1887) for the study of human response to a physical stimulus in a quantitative fashion. It mainly focus on the relationship between perception of object's attributes, like height, weight, and their physical "real" properties.

Weber devised an elegant paradigm where a blindfolded individual had to detect when he felt that an object of gradually increasing weight was really different.

It was found that this *smallest* or *just noticeable difference* (JND) was proportional to the starting value of the weight.

Weber's law (Con't)

If the mass is doubled, the threshold called smallest noticeable difference also doubles. Such a relation between one's perception and the physical attributes of an object is best described by the following equation:

$$\partial p = k\frac{\partial S}{S}, \tag{1}$$

where ∂p is the differential change in perception and ∂S is the differential increase in the stimulus, for a given stimulus S. One usually solve for k by an appropriate experimental design.

Weber's Law is not always true, but it is good as a baseline to compare performance and as a rule-of-thumb.

Weber's law (Con't)

As can be seen from Equation (1), this yields a logarithmic relationship between stimulus real properties and their perception by humans. Indeed, from the integration of (1), which gives $p = k\ln S + c_0$ where $c_0 = -k\ln S_0$ (when $p = 0$, i.e. no perception at all), we see that S_0 is the stimulus threshold below which subject perceives nothing.

Hence, we have:

$$p = k\ln\frac{S}{S_0}, \tag{2}$$

which reads: If a stimulus varies as a geometric progression (i.e. multiplied by a fixed factor), the corresponding perception is altered in an arithmetic progression (i.e. in additive constant amounts).

Weber's law: Illustration

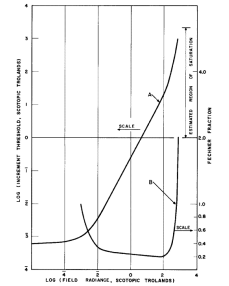


Fig. 3(7.5). Approach to saturation of rod mechanism in 9° extrafoveal vision by rapid rise in increment threshold. *Curve A*: log (increment threshold); *Curve B*: ratio (increment threshold/ field radiance). (Aguilar and Stiles, 1954).

The same relation applies to other stimulus attributes, e.g. light intensity. The ratio of the increment threshold, ΔI, to the background intensity, I, is a constant (k).
If you measure increment thresholds on various intensity backgrounds, you will observe that the thresholds increase in proportion to the background.

Data from Aguilar and Stiles [1].

Stevens's law

Stevens (1906–1973) extends this relationship between stimulus perceived attributes and its real properties,

In its general form, it is formulated as

$$\Psi(I) = kI^a \tag{3}$$

Other well-known laws

More recent work points to

[9, 10, 11]

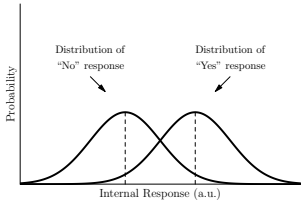
But see [5] in *Rasch Measurement Transactions* (online).

Signal Detection Theory

Signal Detection Theory (SDT) provides a general modeling framework for individual decision taken under uncertainty. It has mostly found application in the study of perceptual and motor skills [2].

As a threshold model it provides a useful introduction to most elaborated probabilistic models that we will introduce in Chapters 7 and 8 (IRT modeling).

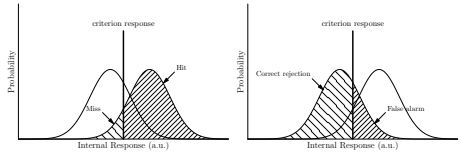
Probabilistic response curves



The above picture describes the distribution of internal response probability for negative (or *noise*) and positive (or *signal-plus-noise*) trials or responses.

Probabilistic response curves

Let an individual be facing a two-alternative choice experiment as before. Depending on the location of his internal criterion, his response may lead to Hit or False Alarm (response > criterion), or alternatively Correct Rejection or Miss (response < criterion).



Probabilistic response curves

In fact, SDT provides a gentle introduction to the statistical decision theory developped by Neyman & Pearson [13], but see [7], which aims at finding good rules for choosing from a specified set of possible actions.

> . . . any rule R prescribing that we take action A when the sample point . . . falls within a specified category of points, and that we take action B in all other cases, is a test of a statistical hypothesis.
>
> Neyman, [12, p. 258]

blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla blabla.

Probabilistic response curves

Most classical textbooks on Statistics provide a Table similar to the one below, where we describe the probabilities of incorrectly rejecting a null hypothesis (α) *vs.* falsely "accepting"[1] the null (β) where in fact the alternative is true.

		Reality	
		H_0	H_1
Test	H_0	$1-\alpha$ (TN)	β (FN)
	H_1	α (FP)	$1-\beta$ (TP)

[1] We never accept the null hypothesis: We just cannot reject it.

Computing discriminability

In the preceding table, TP and TN are both correct decisions (positive and negative response, resp.): They are called 'Hit' and 'Correct rejection', or CR. In the wrong cases, we talked of 'Miss' (FP) and 'False alarm', or FA (FN in the Table).

Given the constraints on the fixed margins, we see that we basically need the two following independent statistics:

- *Hit rate* = Hits/(Hits+Misses), where Hits+Misses is FP+TP (i.e. correct positive decision among all positive trials);
- *FA rate* = FA/(FA+CR), where FA+CR is TN+FN (i.e. incorrect negative decision among all negative trials).

Computing discriminability: Example

C. Pallier [14] gently provides R functions to estimate D' and A' (but see `http://www.pallier.org//ressources/tpexp2/tp.html` for an online tutorial in psycholinguistic). Also, the `psyphy` package provides D' computation for various experimental design (ABX, AFC, . . .).

Psychometrics: Some 'landmarks'

Most significant contributions: *Psychometrische Untersuchungen* (1886), *Essentials of Psychological Testing* (1949), *Probabilistic models for some intelligence and attainment tests* (1960/1980), *Statistical Theories of Mental Test Scores* (1968). . .

F. Galton (1822–1911) K. Pearson (1857–1936) L. L. Thurstone (1887–1955) L. J. Cronbach (1916–2001) A. Jensen (1923–)

J. McKeen Cattell (1860–1944) C. Spearman (1863–1945) G. Rasch (1901–1980) F. M. Lord (1912–2000)

Complete bibliographies may be found at `http://psychclassics.yorku.ca/`.

Why are we measuring cognitive traits?

High-level human functions are more "delicate" to measure. Possible reasons include: less formalized or understood processes (e.g. Mind theory), subject to attrition effect, more subjective in essence, etc.

> Measurement in psychology and physics are in no sense different. Physicists can measure when they can find the operations by which they may meet the necessary criteria; psychologists have but to do the same. They need not worry about the mysterious differences between the meaning of measurement in the two sciences.
>
> Reese, 1943, p. 49

A long standing tradition in Psychology

Much of the current personality and intelligence testing in clinical psychology date back to the late Victorian psychology. This was held under the appelation of "psychology of individual differences" and was promoted by F. Galton and K. Pearson, followed by psychometric testing at University College London with Spearman, Burt and Eysenck.

It is worth noting that such mainstream in "differential psychology" emanates in the context of World War II, the Nazi period, but especially eugenics. . .

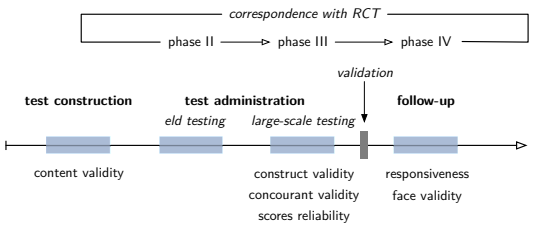
What is a measurement scale?

The classification proposed by Stevens [16] is widely used, although still a matter of debate. Following [15], a variable is said to be

- Ordinal if $f(A) > f(B) \Leftrightarrow A > B$
- Interval if $f(A - B) > f(C - D) \Leftrightarrow (A - B) > (C - D)$
- Ratio if $f(A/B) > f(C/D) \Leftrightarrow A/B > C/D$ (the stronger one)

This has implications in measurement because if we observe interactions at the level of observed scores, does this mean that there are interactions between latent scores? *Yes*, if we have interval metrics or better.

Questionnaire development



Items writing: Some guidelines

Not much to say about item generation, see ???. Some rule of thumbs:

- Ask items with direct content relevance (but also theoretical relevance, i.e. avoid circular assertions);
- Ask items that discriminate known groups;
- Choose items that are maximally independent and that have highest validities;
- Select items to represent single domain.

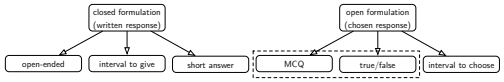
Items writing: Some guidelines

We seek to measure proficiency on a dimension (or *construct*) not observable directly, and we need an instrument of measurement allowing to quantify individual performance in a standardized manner: This is a *normative approach*.

> A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as an knowledge, ability, predisposition, or trait) may be infered.
>
> Osterlind (1990, p. 3)

Items writing: Some guidelines

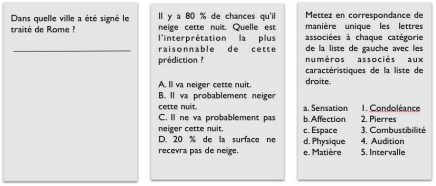
There are several types of items that can be used, as seen below:



Providing the following conditions are fullfilled: (1) the question is clear, (2) the answer is short, (3) marking is easy and simple, MCQs contribute to enhance validity, reliability and sensibility of an exam or a test. MCQs are useful proxys for evaluating performance during training or for evaluating purpose (e.g. grading in a educational context).

Examples of items

Below are general-purpose items (in french) for assessing one's 'culture'.



Dans quelle ville a été signé le traité de Rome ?

_____

Il y a 80 % de chances qu'il neige cette nuit. Quelle est l'interprétation la plus raisonnable de cette prédiction ?

A. Il va neiger cette nuit.
B. Il va probablement neiger cette nuit.
C. Il ne va probablement pas neiger cette nuit.
D. 20 % de la surface ne recevra pas de neige.

Mettez en correspondance de manière unique les lettres associées à chaque catégorie de la liste de gauche avec les numéros associés aux caractéristiques de la liste de droite.

a. Sensation	1. Condoléance
b. Affection	2. Pierres
c. Espace	3. Combustibilité
d. Physique	4. Audition
e. Matière	5. Intervalle

(a) (b) (c)

The MOS SF-36 questionnaire (excerpt)

Below are nine items from the MOS 36 items short form; each item is scored 1 ("All of the Time") to 6 ("None of the Time") points.

		1	*2*	*3*	*4*	*5*	*6*
23.	Did you feel full of pep?	□	□	□	□	□	□
24.	Have you been a very nervous person?	□	□	□	□	□	□
25.	Have you felt so down in the dumps that nothing could cheer you up?	□	□	□	□	□	□
26.	Have you felt calm and peaceful?	□	□	□	□	□	□
27.	Did you have a lot of energy?	□	□	□	□	□	□
28.	Have you felt downhearted and blue?	□	□	□	□	□	□
29.	Did you feel worn out?	□	□	□	□	□	□
30.	Have you been a happy person?	□	□	□	□	□	□
31.	Did you feel tired?	□	□	□	□	□	□

The item seen as an instrumental variable

An item can be considered as a criteried way to measure a psychological construct otherwise unobservable (e.g. reading fluency, specific cognitive ability). Most importantly, questions must be *decidable*.

How to define an MCQ?
This is a question where a subject has to operate a selection and give (at least) one answer among different possible responses, each being either correct or not independently of the person to whom it was submitted to (Leclercq (2000, p. 15)). The correctness of the answer(s) is defined by external entities (test constructor and expert consensus).

More about MCQs

In a certain sense, MCQs are a variant of questions with dichotomous responses. They have comparable reliability and concourant validity [6]. Of special importance is the fact they are not necessarily less reliable as usually thought.

For a test with 5 'true'/'false' items, the probability that a subject responding at random gives 5 correct answers is $\frac{1}{2^5} = 0.03$. Mathematical expectation is $5 \times 0.5 = 2.5$. For a test composed of 5 items with 4 aternative responses, the same corresponding probability is $\frac{1}{4^5} = 0.00098$ (and $\mathbb{E}(X_1 = 1, X_2 = 1, \ldots, X_5 = 1) = 1.25$).

How well do binary items perform?

According to [8], the following factors may influence observed performance:

- item keying
- item formulation (positive *vs.* negative)
- truthfulness of a solution with respect to its attractiveness
- effect of negative wording of the proposed answer

Influence of item keying

Consider the next two cases: Subject has to independently evaluate the proposed answers *vs.* subject knows in advance how many response are correct. Would you predict that he will perform better in the second situation?

In the first case, this is an identification task whereas in the second case we talk of a discrimination task [3, p. 171].

Influence of item keying

Manipulating two kind of assertion—positive or negative—Fabre & Noizet [4] observed a slight asymmetry in favor of true negative assertion compared to false negative: This reflects a general tendency to *acquiescence*.

	Solution		
Formulation	True	False	
Affirmative	83.3	50.7	67.0
Negative	52.3	76.7	64.5
	67.8	63.7	

Item analysis

In the vein of Classical Test Theory (more on CTT in the next chapter), the quality of a questionnaire depends in part on the psychometrical properties of items it is composed of. The same applies for surveying techniques in general.

Item analysis mainly focuses on (1) item parameter or location on the latent trait, and (2) discriminatory power, whereas at the test level we are interested in (3) internal consistency and (4) scores reliability, or standard error of measurement.

Item difficulty

Item parameter reflects the location of the item on the underlying latent trait. In educational assessment, this is referred to as *item difficulty*. It is generally estimated by the proportion of correct responses, and is (unfortunately) called the p-value or observed p of the item. To avoid any confusion, we will denote it as p_{obs}.

An item is considered difficult when its corresponding p_{obs} is low. However, this statistic depends on the sample under consideration. For instance, if we have only high-graded students, all items would have low difficulty, whereas with less competent students the same items would be judged more difficult.

Item difficulty: Illustration

On the right, we see the p_{obs} of five items. It is clear that item 4 is the most difficult (36% of the subjects gave a correct answer). But this is only true for this particular set of subjects!

Item	p-obs
1	.68
2	.45
3	.91
4	.36
5	.48

Item functionning

If we want to get a better idea of how a multiple choice item is behaving, we can look at the distribution of all k response categories (if it is an MCQ-like item, there is only one key and $k-1$ distractors, $k>2$).

	A	B	C*	D	NR	Total
N	28	17	197	41	3	286
p-obs	.10	.06	.69	.14		

In the above Table, the correct answer (key) is 'C' and it is chosen by 69% of the subjects. The other response category are equally chosen in about 10% of the cases.

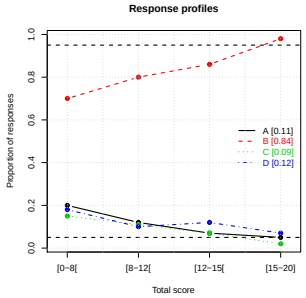
Item functionning: Illustration

The following two Tables illustrates dysfunctionning items. In the left panel, we see that there is a problem with distractor 'C' which is chosen more often than the key 'A', and distractor 'B' is never used; on the right, the key is not answered by the majority of the subjects (this is probably caused by a bad association when scoring the item).

	A*	B	C	D	NR	Total
N	77	0	130	63	16	286
p-obs	.27	**.00**	.45	.22		

	A	B*	C	D	NR	Total
N	202	31	28	25	0	286
p-obs	**.71**	.11	.10	.9		

Item functionning: Illustration (Con't)



References

[1] M Aguilar and W S Stiles. Saturation of the rod mechanism of the reina at high levels of stimulation. *Optica Acta*, 1:59–65, 1954.

[2] Claude Bonnet. *Manuel pratique de psychophysique*. Armand Colin, 1986.

[3] J M Fabre and G Noizet. Confiance attachée aux réponses à des questions à choix multiple. *Journal de Psychologie normale et pathologique*, 74:355–362, 1977.

[4] J M Fabre and G Noizet. Jugement du vrai et du faux sur des énoncés affirmatifs et négatifs. In *Psychologie expérimentale et comparée, Hommage à Paul Fraisse*, volume 74, pages 411–426. Paris, P.U.F., 1977.

[5] W P Fisher. Ordinal vs. ratio revisited again. *Rasch Measurement Transactions*, 18(2):980–982, 2004. `http://www.rasch.org/rmt/rmt182e.htm`.

[6] D A Frisbie. Multiple-choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement*, 10:297–304, 1973.

[7] Ian Hacking. *Logic of Statistical Inference*. Cambridge University Press, 1965.

[8] D Leclercq. *La conception des questions à choix multiple*. Bruxelles, Labor, 2000.

[9] R D Luce. On the possible psychophysical laws. *Psychological Review*, 66:81–95, 1959.

[10] R D Luce. "On the possible psychophysical laws" revisited: Remarks on cross-modal matching. *Psychological Review*, 97:66–77, 1990.

[11] R D Luce. A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review*, 109(3):520–532, 2002.

[12] J Neyman. *First Course in Probability and Statistics*. New York: Holt, 1950.

[13] J Neyman and E S Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, A*, CCXXXI:289–337, 1933.

[14] C Pallier. Computing discriminability and bias with the R software, 2002. Available at: `http://www.pallier.org//ressources/aprime/aprime.pdf`. Accessed March 1, 2010.

[15] W Revelle. *An introduction to psychometric theory with applications in R*. 2010. `http://www.personalitytheory.org/r/book/`.

[16] S S Stevens. *Psychophysics*. Transaction Books, 2000.

– Typeset with FoilTEX (version 2)