

USER, COME BACK!!

Maxine Eskenazi

Language Technologies Institute, CMU

June 16, 1999

What is the goal?

**Carnegie
Mellon**

Assessment differs when the application becomes “real”

Then you can based it on the volume of calls

As long as we do not make “real” reservations, is user satisfaction valid?

Users can point to places (“hotspots”) where they had a hard time

What do they remember that would make them call or not call again?

We would want the system with the least “hotspots”

System “hotspots”

**Carnegie
Mellon**

What do you remember about the last system you called?

“It kept repeating the information about one hotel.”

Why was it irritating?

“I had to repeat myself”

“The system asked the same thing over and over”

“I told it Pittsburgh and then it asked where I wanted to go”

“It wasn’t talking about the same thing I was”

Measures

Carnegie
Mellon

label turns by subgoal (departure time, hotel location)

sr=no. syst repeats; ur=no. user repeats;

T=no. turns; us=user then syst said same thing;

sn=syst subgoal /= user subgoal

repeats (R) = $sr+ur/2T-2$ (per goal)

“I had to repeat myself” , “The system asked the same thing over and over” :

“shadows” (S) = $us/(T-1)$ (per goal)

“I told it Pittsburgh and then it asked where I wanted to go”

system/=user (N) = sn/T

“It wasn’t talking about the same thing I was”

System Assessment

Carnegie
Mellon

Compare semi-automatic measures to what users say

Panel of 9 callers

9 different scenarios, 3x3 different difficulty levels

easy = 1 leg; med. = 2 legs+; hard = 3 legs+

each caller did: 1 easy, 1 medium, 1 hard

Repetitions - questions

**Carnegie
Mellon**

Tell me where in the dialogue:

Something didn't go as you thought it should

You had to change what you wanted

You wanted to give up

You had no idea what to do next

Comparative results

**Carnegie
Mellon**

For 20 dialogues, user/label agreement:

highest number of turns: 65%

highest percent of repeats (R): 85%

highest percent of system “shadows (S): 60%

highest percent of system/=user (N): 55%

(S) seems to refine the information that is in (R)

divides problem repeats from simple navigation

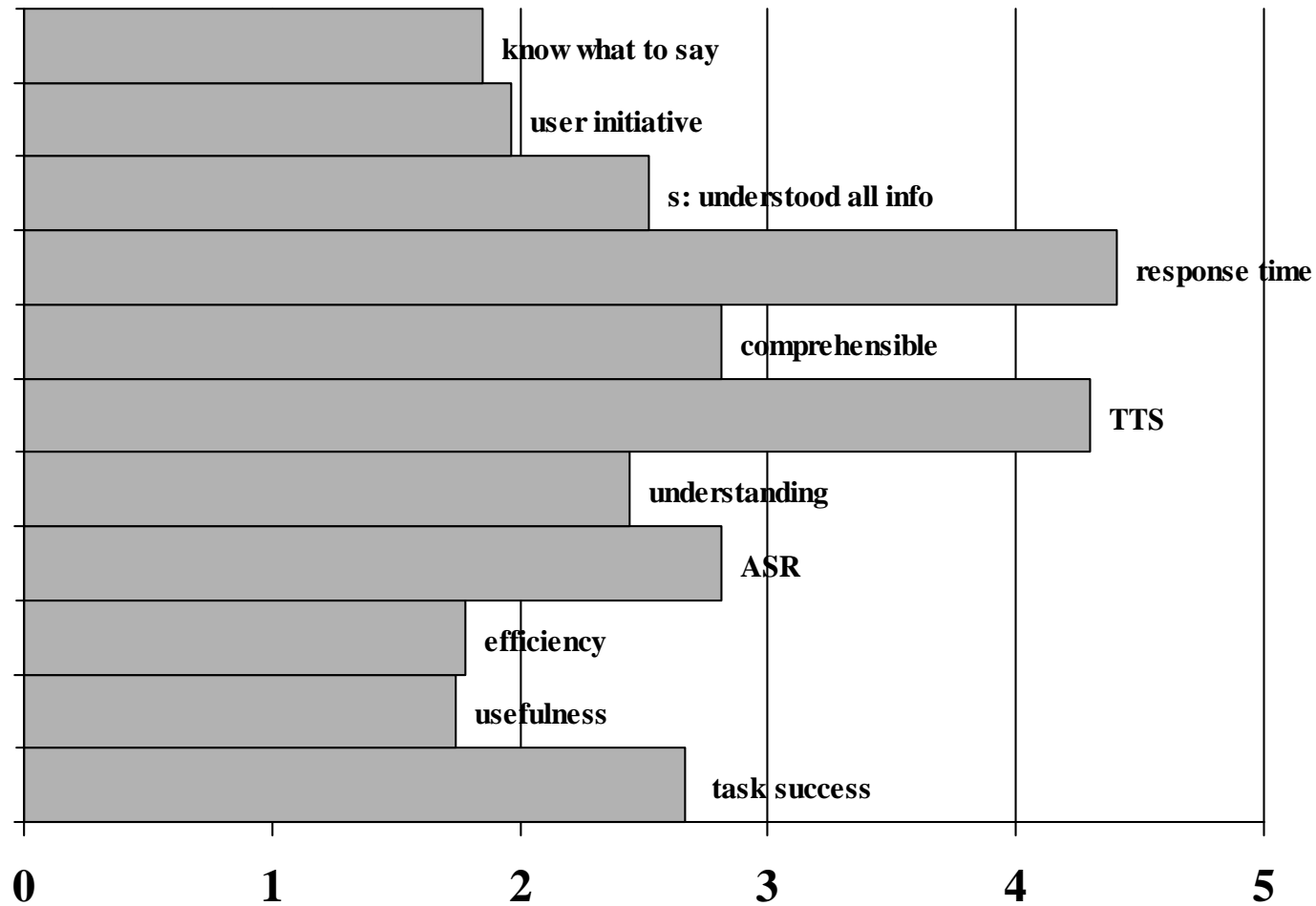
Comparative Results - Task Difficulty

**Carnegie
Mellon**

	Turns	Repetitions	Shadows	S/=U
All	65%	85%	60%	55%
“Easy”	50%	100%	100%	50%
“Medium”	62.5%	75%	62.5%	75%
“Hard”	80%	40%	40%	20%

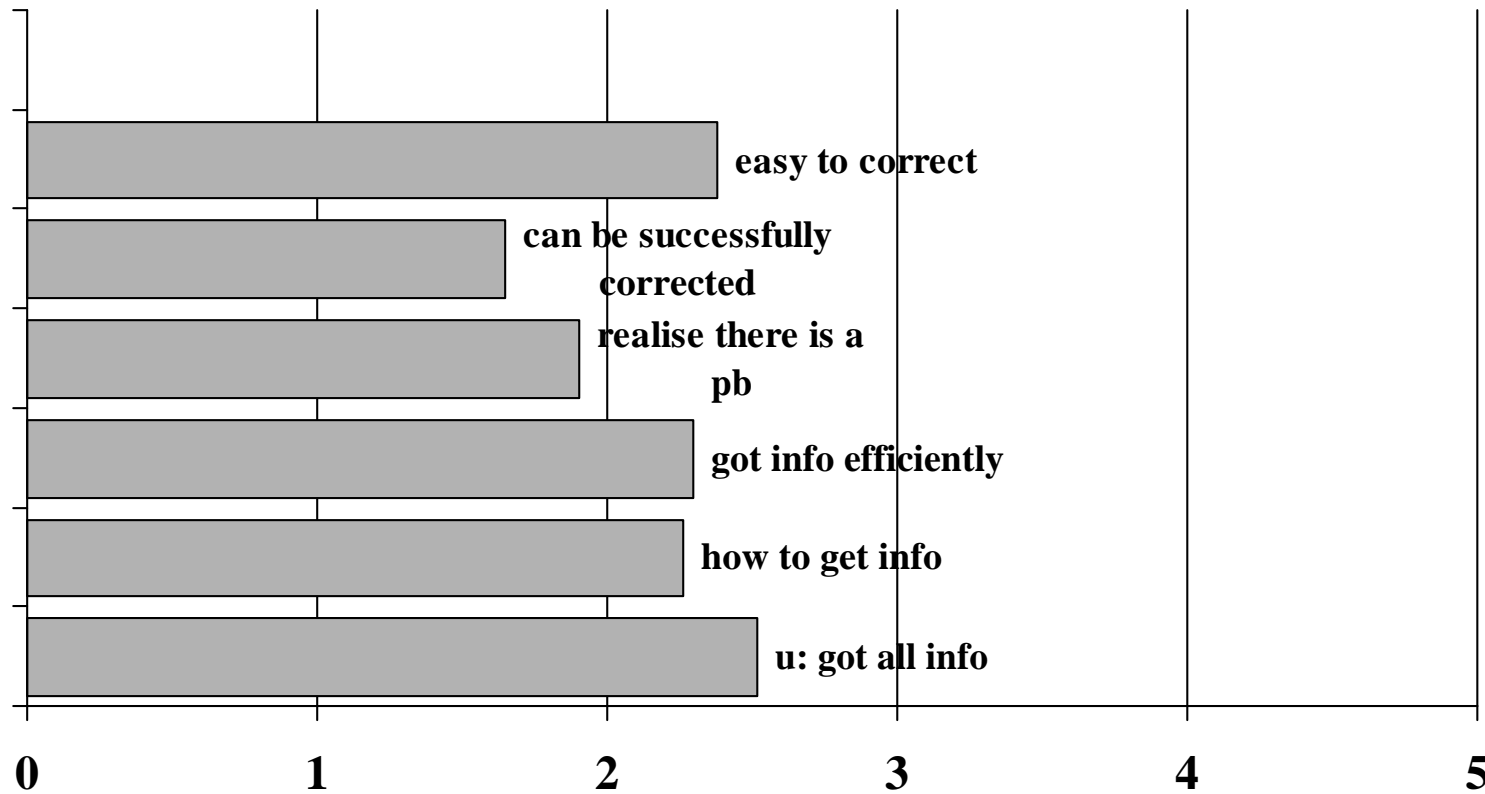
Assessment - Sanders questionnaire

Carnegie
Mellon



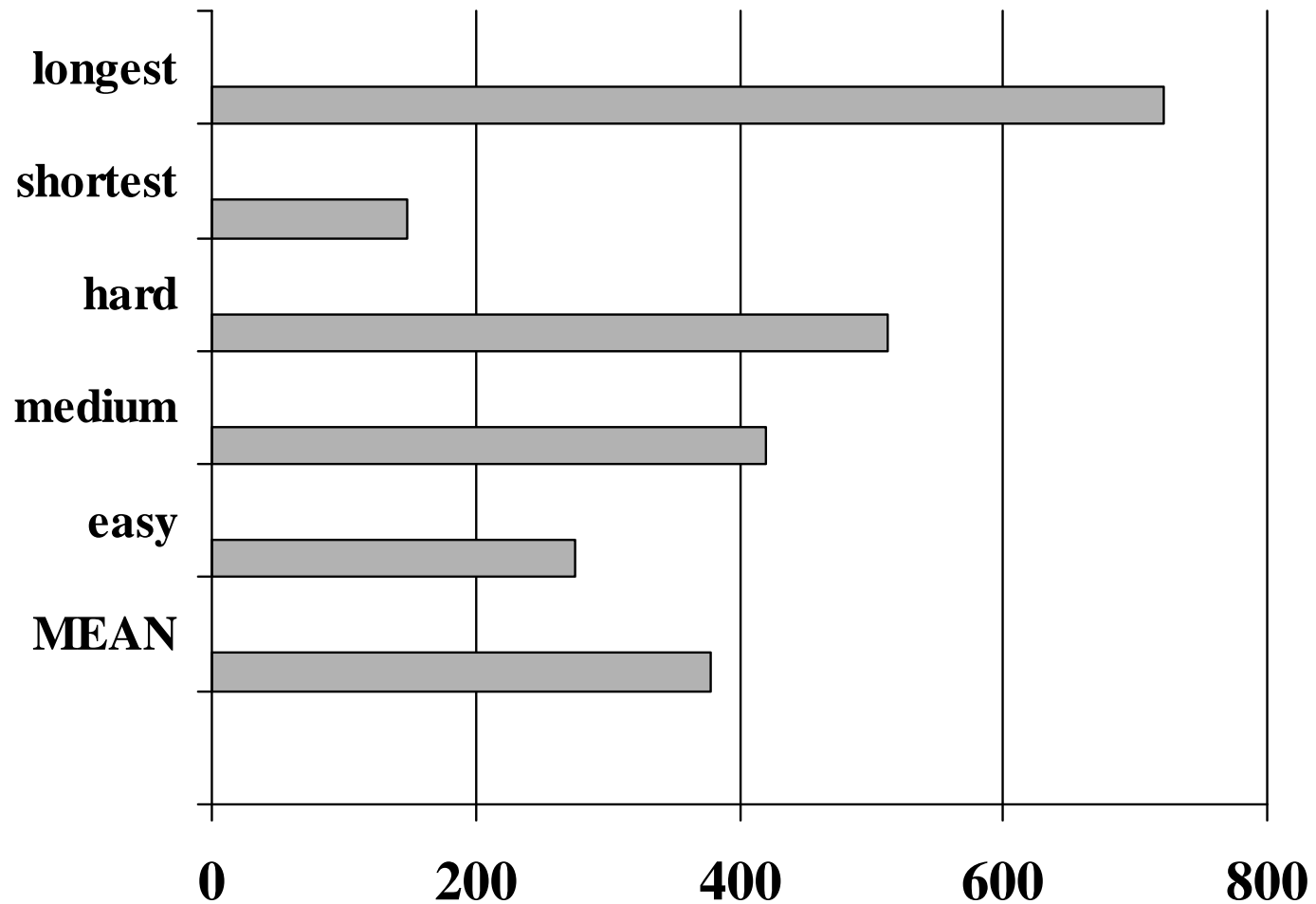
Sanders questionnaire - cont.

**Carnegie
Mellon**



Length of dialogues

Carnegie
Mellon



Acquired Data

**Carnegie
Mellon**

<u>Type</u>	<u>No. Dialogs</u>	<u>No. Utts.</u>	<u>Time</u>
Human-Human	58	1800	~1 hr
WOZ1	107	1992	~1.3 hrs
WOZ2	16	487	~20 min
Movieline	112	7527	~3 hrs
System	2861	44783	~11.4 hrs

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.