CHAPTER

39

Stimulus-Driven Guidance of Visual Attention in Natural Scenes

Derrick J. Parkhurst and Ernst Niebur

ABSTRACT

A large body of research indicates that the focus of attention and eye movements are guided by bottom-up, stimulus-driven mechanisms of visual attention. This research has primarily used simple experimental tasks and simple visual displays in order to maximize experimental control. In this chapter, we discuss recent work that reexamines bottom-up guidance of attention by measuring eye movements made by observers viewing complex, natural scenes. The results of this research indicate that under natural viewing conditions, attention is indeed guided by stimulus-driven mechanisms.

I. INTRODUCTION

The primate visual system receives an enormous amount of information as input and, rather than attempting to fully process all this information, portions of the input are selected for detailed processing while the remaining information is left relatively unprocessed. Two classes of attentional mechanisms control this selection process. Bottom-up selection involves fast, and often compulsory, stimulus-driven mechanisms. That is to say, computational resources are allocated to particular parts of the visual input, based on the properties of that input. For example, attention is preferentially allocated to unique features, abrupt onsets, and the appearance of new perceptual objects. Top-down selection is typically slower and governed by the observer's expectations, intentions, or memory. For example, observers can volitionally select objects or regions of space for detailed processing. Note, however, that intentionality is not a necessary component of top-down selection because familiar scene contexts (Chapter 40) and semantic associatiotis (Moores *et al.*, 2003) can influence attention, even in the absence of the explicit knowledge of the observer.

The majority of the research on visual attention has used relatively simple experimental paradigms designed to obtain a high degree of experimental control. Simplified visual stimuli, for example, visual search arrays consisting of colored bars of varied orientations, are nearly always used in conjunction with visual search tasks (see Chapter 17). This research has been extremely valuable. However, it is not clear whether the principles of attentional guidance gleaned from this research generalize to more complex stimuli. Thus, the results obtained should be validated using paradigms that use natural scenes and natural tasks.

The traditional measures of attention (e.g., those inferred through reaction times or error rates) are not easily determined for natural viewing conditions. However, important insights into the allocation of attention can be derived by examining the way in which peop]e make eye movements. The logic of this approach rests on the assumption that eye movements and attention are correlated. This assumption is a reasonable one given that both eye movements and attention are related to the selection of the most important parts of the visual input. Although the locations of the focus of attention and the center of gaze can be dissociated, psychophysical evidence indicates that focal attention at the location of a pending eye movement is a necessary precursor for that movement (for review, see Chapter 20).

II. EYE MOVEMENTS IN NATURAL SCENES

Over the years, a number of studies have examined eye movements recorded from observers' viewing complex natural scenes and doing a variety of dif-

ferent tasks. For the most part, these studies have made qualitative claims about the relationship between eye movements and stimulus properties. Some of the earliest of these studies indicated that observers preferentially look at people and faces, although this result depended heavily on the task of the observers (Yarbus, 1967). Later, more quantitative analyses indicated that observers look at regions that are deemed to be informative (Antes, 1976). It is only recently that extensive quantitative analyses have begun to be used to examine the relationship between stimulus properties and eye movements. This approach has become feasible given rapidly improving eye tracking techniques and readily available computational resources for image processing. The majority of these quantitative studies have shown a significant correlation between stimulus features and eye movements (Mannan *et al.*, 1996; Reinagel and Zador, 1999; Krieger *et al.*, 2000) in free-viewing paradigms. This suggests that image features guide attention in a bottom-up fashion under natural conditions.

To further quantify the relationship between stimulus features and eye movements, we recently recorded eye movements from participants viewing a variety of natural and artificial scenes including home interiors, fractals, natural landscapes, and city scenes (Parkhurst and Niebur, 2003). Participants were told to free-view the images and that the only requirement was that they look around in the images. Presented in Fig. 39.1 are examples of the natural landscapes used in the experiment accompanied by the quantitative results of our eye movement analyses. We focus on the results obtained using the natural landscapes database because they are characteristic of the pattern of results obtained with other image databases.

We began by examining the relationship between local contrast and the observed fixation locations. To accomplish this, image patches were extracted from the images at the observed fixation locations and contrast was calculated as the standard deviation of pixel intensities within each patch. We refer to this ensemble of extracted images patches as the participant-selected ensemble. If contrast attracts attention, we expect the contrast in the participant-selected ensemble to be greater than that expected by chance factors alone. To test this prediction, we first estimated the contrast expected by chance using the average contrast in each of two additional image ensembles, a uniformly selected ensemble and an image-shuffled ensemble. The uniformly selected ensemble was created by extracting patches from random locations in each image. The contrast obtained using this ensemble serves as an estimate of the average image statistics of the images. The image-shuffled ensemble is used to control for the fact that participants may not sample the images uniformly; for instance, subjects typically show a bias to fixate central locations. This ensemble is created by extracting image patches at the observed fixation locations (tending to be central) but using a shuffled image database. Although this procedure equates the distribution of fixation locations in the participant-selected and image-shuffled databases, using it as a baseline comparison may be overly conservative given that the central fixation bias is likely to be due, at least in part, to a greater presence of interesting stimulus features near the center of the images.

As can be seen in Fig. 39.1B, the average contrast for the participant-selected ensemble (dashed line, circle) is always significantly greater than that obtained with the uniformly selected ensemble (solid line, square) or the image-shuffled ensemble (solid line, triangle). This result indicates that regions of high contrast tend to be fixated under natural viewing conditions and suggests that attention is guided to regions of high contrast in a bottom-up fashion.

The question of why attention should be drawn to regions of high contrast presents itself. One answer is that these regions may tend to be more informative for accomplishing behaviorally relevant tasks, for example, searching for and recognizing objects in a natural scene. A simple, purely stimulus-based, measure of the informativeness of a region is the correlation between local pixel intensities. Although it is well known that local intensities in images of natural scenes tend to be correlated due to common lighting, the degree of correlation can vary dramatically across different locations in a scene. For example, correlation will be low for regions that contain luminance discontinuities, such as edges, whereas the correlation is high for uniform regions, such as surfaces. Note that this measure of correlation differs from a measure of contrast in that the structure of the image patch is important. Whereas low contrast necessarily implies a high correlation between the intensity at different locations, high contrast can cooccur both with low correlation (e.g., with a random noise stimulus) or with high correlation (e.g., with a sine wave, a checkerboard Pattern, or more complex patterns).

To explore the dependence of eye movements on the structure of the scene, we used the two-point correlation function. It is defined between the points at the center of each patch (i.e., at the observed fixation locations) and the points at a given distance from the center. The correlations obtained for each of the three image patch ensembles are shown in Fig. 39.1C. As expected for natural scenes, the correlations are highest for short distances and monotonically decrease with increasing distances. More important, the corre-

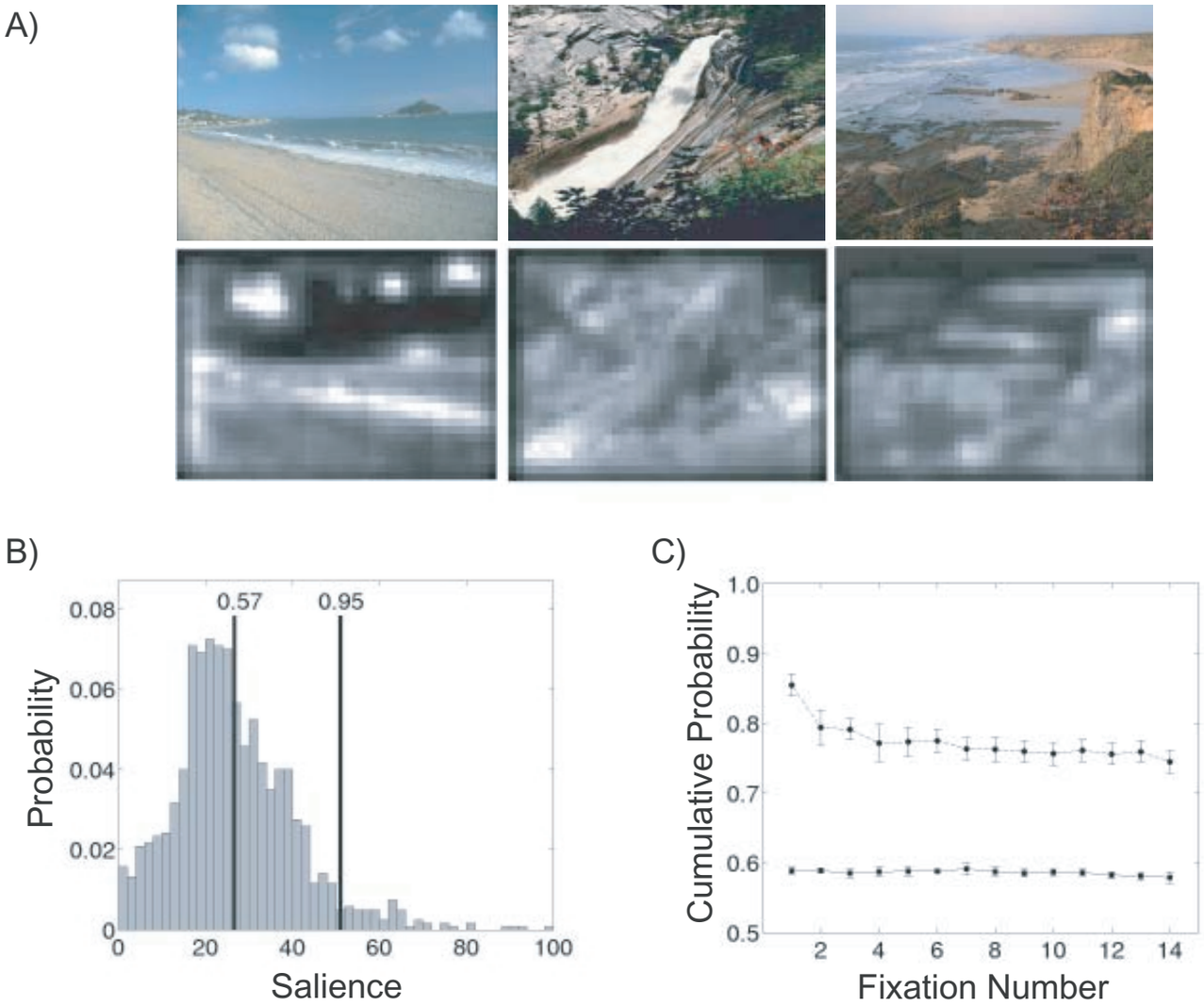

FIGURE 39.1 (A) Example natural landscapes. (B) Average contrast in the participant-selected ensemble (dashed line, circle), uniformly selected ensemble (solid line, square) and image-shuffled ensemble (solid line, triangle) all as a function of image patch size. (C) Two-point correlation using a 4-deg radius patch as a function of the distance from fixation. Same symbols as in (B). Error bars represent plus/minus one standard error of the mean taken across random permutations.

lations observed with the participant-selected image ensemble tends to be less than those observed for the uniformly selected or image-shuffled ensembles. This indicates that it is not regions of high contrast *per se* that attract fixation but rather regions that show an especially low correlation. Regions of low correlation have the highest information content and thus fixating these regions is an ideal strategy to gain information about the stimulus. These results suggest that bottom-up attention guides eye movements in order to maximize information about the stimulus.

III. STIMULUS SALIENCE IN NATURAL SCENES

A number of stimulus properties in addition to luminance attract attention in a bottom-up fashion including color, orientation, and motion. Under controlled conditions, the conspicuousness of a stimulus, also referred to as its salience, depends strongly on local feature contrast (see Chapter 38). For example, consider a red circle on a neutral gray background. The circle is salient because it is the only object in the

display. However, surround this circle with a number of identical red circles and the salience of the original circle is greatly reduced. If we instead surround the red circle by green circles, its salience is greatly enhanced. The salience of a stimulus is related to the contrast between its features and the features of its neighbors. Functionally speaking, salience is determined by the visual uniqueness of a stimulus within the context of a scene, and local feature contrast is one plausible way to calculate the degree of uniqueness of a stimulus.

The question arises whether this analysis of salience, derived from simple displays, generalizes to natural scenes. For example, if the circles are, instead, red and green apples on display at a local market, will attention still be a function of stimulus salience? To answer this question, we might consider investigating the ability of a variety of different stimulus features to guide attention, as we did for luminance contrast in the previous section. However, given the ongoing disagreement over the fundamental set of stimulus features that attract attention in a bottom-up fashion (see Chapter 17), this approach is difficult and at best tedious. In lieu of such an approach, we decided to implement a biologically motivated computational model of the primate visual system and use this model to quantify stimulus salience (Parkhurst *et al.*, 2002). The design of this model, its representations and algorithms, were based on what has been learned about the primate visual system from a large number of neurophysiological and neuroanatomical studies (see Niebur and Koch, 1996; Itti *et al.*, 1998; Parkhurst, 2002, for more details about the model). In this way, the representation of stimulus salience is derived from a single, neurally plausible implementation rather than from a potentially large battery of psychophysical studies. This computational implementation allows us to explicitly quantify stimulus salience and predict attentional allocation in complex natural scenes, our primary interest.

The model takes a photograph of a natural scene as input and processes it in three parallel feature channels, representing luminance, color, and orientation across a range of spatial scales. This processing results in a set of topographic center-surround feature maps. To derive an estimate of stimulus salience, these feature maps are combined across scales and feature channels to form a saliency map (Koch and Ullman, 1985). The saliency map indicates the most salient, or visually unique, regions in the scene. A number of natural scenes and their respective salience maps are shown in Fig. 39.2A.

We reasoned that if attention is indeed a function of stimulus salience under natural viewing conditions, there should be a correspondence between fixation locations and the salience of the stimuli at those locations. To test this logic, we examined the stimulus salience, as determined by our model, at the fixation locations observed in the free-viewing paradigm described in the previous section. We used the following procedure to quantify the correspondence between stimulus salience and fixations. First, the salience at each fixation location was extracted from the relevant salience map and compared to the overall distribution of salience in that map. Note that the distribution of salience in any given salience map is often positively skewed. This is because there can be only a small number of very salient locations in a scene; otherwise these locations, being no longer unique, would no longer be salient. An example distribution is shown in Fig. 39.2B. Next, the probability of finding a salience value less than or equal to that extracted from the observed fixation location was calculated. This is a cumulative probabihiy and is equal to 1.0 if the location of maximum salience is fixated. The cumulative probability expected by chance factors alone, in other words the value that is expected if fixation locations are chosen at random, is the cumulative probability for the average salience value. Example cumulative probabilities for a single fixation are shown as dark bars in Fig. 39.2B. Given the positive skew of the salience distributions, the cumulative probability expected by chance factors alone is approximately 0.6, on average.

We found that the average cumulative probability for the observed fixations significantly exceeded the cumulative probabilities expected by chance. This is shown in Fig. 39.2C, which indicates that there is a significant correspondence between fixation locations and stimulus salience. We also found that the largest effect is seen for fixations made just after stimulus onset (Koch and Ullman, 1985). This is consistent with the time course of top-down attentional mechanisms that are known to have a slower onset than bottom-up mechanisms. These results support the conclusion that attention is indeed guided by bottom-up mechanisms nuder natural viewing conditions. Furthermore, given the magnitude of the effect, bottom-up mechanisms can play an important role in determining the guidance of attention.

IV. CONCLUSION

In order to deal with the complexity of natural scenes, the visual system must select a small portion of the scene to process in detail, leaving the remainder of the scene for processing at a later time, or not at all. Both bottom-up attentional mechanisms, which are dependent on the stimulus, and top-down attentional

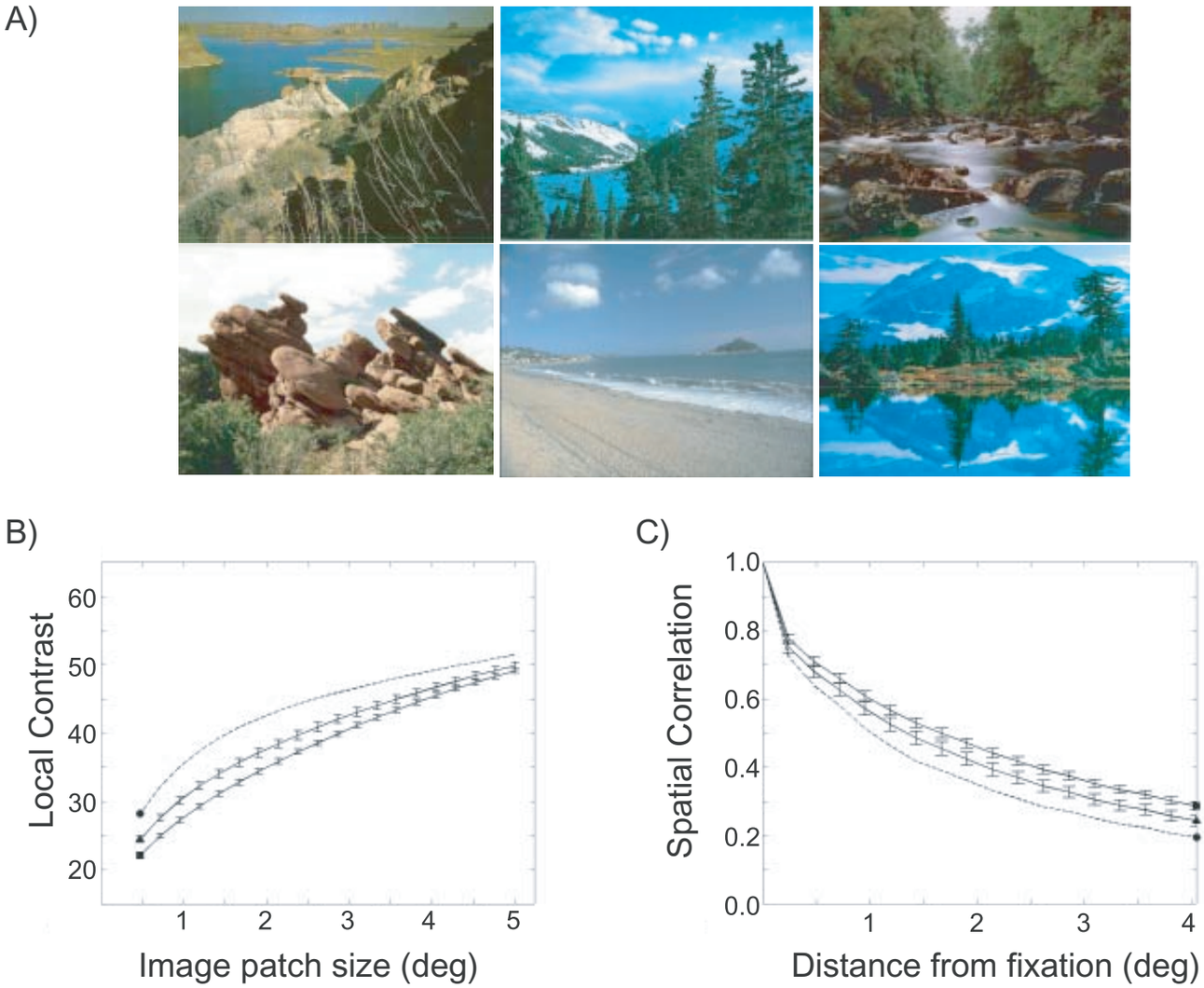

FIGURE 39.2 (A) Example natural landscapes and their respective salience maps. (B) The distribution of salience in an example salience map is shown with the salience expected by chance factors alone (cumulative probability = 0.57) and the salience obtained for a single example fixation (cumulative probability = 0.95). (C) The cumulative probabilities at the points of fixation (dashed line; circle) and expected by chance (solid line; square) are shown as a function of fixation number after stimulus onset. Error bars represent plus/minus one standard error of the mean taken across participants.

mechanismns, which are dependent on the viewer, contribute to this selection process.

In this chapter we described how we quantified the role of bottom-up attentional guidance using eye movements obtained from participants viewing natural scenes. Because these studies were observational rather than experimental in nature, there is the possibility that an unobserved variable not related to stimulus salience could account for the results that we obtained. We argue that this is not likely to be the case given the converging evidence in support of our conclusion from observational, computational, and recent experimental studies (see Parkhurst amid Niebur, 2004).

These studies are just the beginning of our investigation into attentional allocation under natural conditions, and a number of questions remain open. For example, whereas free-viewing of scenes probably captures the way in which we view a scene when we are free from task constraints, how is attentional allocation determined when a complex task needs to be performed? Other open questions include how atten-

tional allocation is affected by episodic memory (e.g., having viewed a scene before; see Chapter 40) and semantic memory (e.g., the gist of a scene; see Chapter 41). More generally, how can the understanding of bottom-up and top-down influences be integrated into a common framework?

We argue that computational modeling of visual processing will be an important tool to answer these questions given the inherent difficulty of studying attentional allocation in natural scenes. A number of important insights into attentional guidance in natural scenes have come from modeling approaches (e.g., see Chapters 65 and 96). Computational models allow for explicit implementations of conceptual hypotheses and can make quantitailve predictions for complex, natural stimuli. However, it is important that the modeling be integrated with behavioral and neuroscientific approaches to achieve its full potential.

Acknowledgments

This work was supported by NSF through a CAREER award to EN. Derrick Parkhurst was also supported by a NIH-NEI postdoctoral training fellowship.

References

Antes, J. R. (1976). The time course of picture viewing. *J. Exper. Psychol.* **103**, 62–70.

Itti, L., Niebur, E., and Koch, C. (1998). A model of saliency-based fast visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **20**, 1254–1259.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiol.* **4**, 219–227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., and Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vis.* **13**, 201–214.

Mannan, S. K., Ruddock, K. H., and Wooding, D. S. (1996). The relationship between the locations of spatial features and those fixations made during visual examination of briefly presented images. *Spatial Vis.* **10**, 165–188.

Moores, E., Laiti, L., and Chelazzi, L. (2003). Associate knowledge controls deployment of visual selective attention. *Nature Neurosci.* **6**, 182–189.

Niebur, E., and Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. *In* "Advances in Neural Information Processing Systems" (D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds.), Vol. 8, pp. 802–808. MIT Press, Cambridge, MA.

Parkhurst, D. (2002). "Selective Attention in Natural Vision: Using Computational Models to Quantify Stimulus-Driven Attentional Allocation." Unpublished Ph.D. iss., Johns Hopkins University, Baltimore, MD.

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual selective attention. *Vis. Res.* **42**, 107–123.

Parkhurst, D. J., and Niebur, E. (2003). Scene content selected by active vision. *Spatial Vis.* **6**, 125–154.

Parkhurst, D. J., and Niebur, E. (2004). Texture contrast attracts overt attention in natural scenes. *Eur. J. Neurosci.* **19**, 783–789.

Reinagel, P., and Zador, A. M. (1999). Natural scene statistics at the center of gaze. *Network: Comput. Neural Syst.* **10**, 341–350.

Yarbus, A. (1967). "Eye Movements and Vision." Plenum Press, New York.

AUTHOR QUERY FORM

Dear Author,
During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.
Many thanks for your assistance.

Query References	Query	Remarks
1	*Au*: Please verify citation	