Acta Psychologica xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

On the efficiency of instruction-based rule encoding

Hannes Ruge*, Tatjana Karcz, Tony Mark, Victoria Martin, Katharina Zwosta, Uta Wolfensteller

Technische Universität Dresden, Department of Psychology, Germany

ARTICLE INFO

Keywords:
Instruction-based learning
Rapid instructed task learning
Trial-and-error learning
Feedback
Working memory
Automatization

ABSTRACT

Instructions have long been considered a highly efficient route to knowledge acquisition especially compared to trial-and-error learning. We aimed at substantiating this claim by identifying boundary conditions for such an efficiency gain, including the influence of active learning intention, repeated instructions, and working memory load and span. Our experimental design allowed us to not only assess how well the instructed stimulus-response (S-R) rules were implemented later on, but also to directly measure prior instruction encoding processes. This revealed that instruction encoding was boosted by an active learning intention which in turn entailed better subsequent rule implementation. As should be expected, instruction-based learning took fewer trials than trial-and-error learning to reach a similar performance level. But more importantly, even when performance was measured relative to the identical number of preceding *correct* implementation trials, this efficiency gain persisted both in accuracy and in speed. This suggests that the naturally greater number of failed attempts in the initial phase of trial-and-error learning also negatively impacted learning in subsequent trials due to the persistence of erroneous memory traces established beforehand. A single instruction trial was sufficient to establish the advantage over trial-and-error learning but repeated instructions were better. Strategic factors and inter-individual differences in WM span – the latter exclusively affecting trial-and-error learning presumably due to the considerably more demanding working memory operations – could reduce or even abolish this advantage, but only in error rates. The same was not true for response time gains suggesting generally more efficient task automatization in instruction-based learning.

1. Introduction

In recent years, an increasing number of behavioral and brain imaging studies have aimed at understanding the cognitive and neural mechanisms by which explicitly instructed novel rules come to exert control over behavior. In contrast to trial-and-error learning where the learner encodes retrospectively – based on performance feedback – whether his or her behavior was appropriate in the given situation, instruction-based learning allows the learner to encode prospectively whether a future behavior will be appropriate if certain situational conditions are met (Cole, Laurent, & Stocco, 2013; Noelle, 1997; Ruge & Wolfensteller, 2016b; Wolfensteller & Ruge, 2012). However, little is known about how instruction encoding processes translate into the successful behavioral implementation of the instructed task later on and, moreover, how this compares to trial-and-error learning. The present paper reports the results of three studies that aimed at elucidating these two open issues. To be able to pursue this goal we used an experimental design that provided behavioral markers of both, the initial instruction encoding process and the subsequent behavioral implementation of the just instructed task in terms of accuracy and speed.

This approach contrasts with a related but also clearly different agenda pursued by several previous studies that have examined the ‘power of instructions’ by assessing how merely instructed behavioral rules of an ‘inducer task’ would interfere with the execution of a subsequent ‘diagnostic task’ which requires behavior according to *other* than the instructed rules (Duncan et al., 2008; Everaert, Theeuwes, Liefooghe, & De Houwer, 2014; Meiran, Pereg, Kessler, Cole, & Braver, 2015; Waszak, Wenke, & Brass, 2008; Wenke & Frensch, 2005). Moreover, these studies typically did not obtain a direct measure of instruction encoding processes to assess how these might be related to their subsequent impact on overt behavior. For instance, in a study by Meiran et al. (2015), the inducer task instruction was to press the left key for one arbitrary symbol and the right key for another arbitrary symbol. Additionally, subjects were generally instructed to implement the inducer task in subsequent trials only if the symbols were displayed in green color. If symbols were displayed in red color, subjects were required to press a fixed response being either compatible or incompatible with the response that was linked to this symbol by the previous inducer task instruction. In the basic version of the paradigm, the

---

[*] Corresponding author at: Technische Universität Dresden, Fakultät Mathematik und Naturwissenschaften, Fachrichtung Psychologie, 01062 Dresden, Germany.
E-mail address: hannes.ruge@tu-dresden.de (H. Ruge).

http://dx.doi.org/10.1016/j.actpsy.2017.04.005
Received 18 October 2016; Received in revised form 11 April 2017; Accepted 11 April 2017

Please cite this article as: Ruge, H., Acta Psychologica (2017), http://dx.doi.org/10.1016/j.actpsy.2017.04.005

diagnostic task started directly after the inducer task instruction. This so-called 'NEXT phase' consisted of a sequence of trials where one of the two symbols was randomly selected and presented in red color. Hence, the impact of the merely instructed inducer task could be measured by the performance difference between compatible trials (i.e., inducer task response the same as the NEXT response) and incompatible trials (i.e., inducer task response different from the NEXT response).

Importantly, the theoretical stance of this kind of studies was more on demonstrating an *unintended*, automatic or reflex-like impact of merely instructed tasks even if the instructed task was completely irrelevant for the correct implementation of the diagnostic task. By contrast, the empirical work presented in this paper aimed at assessing the *intended* impact of an instructed S-R task on the subsequent implementation of this very task. As earlier examples of this general approach, previous brain imaging studies monitored activation during the encoding of instructed stimulus-response (S-R) rules and identified brain regions in which encoding-related brain activation predicted subsequent performance (Demanet et al., 2016; Ruge & Wolfensteller, 2010, 2016a). In the present series of experiments, however, we directly assessed encoding-related processes by means of behavioral measures. Specifically, instead of encoding-related brain activation we determined instruction encoding time indexed by behavioral response latency as a presumed direct measure of how much effort is being invested into the encoding of the instructed rules. We used this experimental setup as a starting point to examine the efficiency of instruction-based rule encoding processes and sought to identify possible boundary conditions in general and specifically regarding an often presumed efficiency gain relative to feedback-driven trial-and-error learning.

Study 1 examined the relevance of intentional or 'active' instruction encoding as compared to physically identical but psychologically rather incidental or 'passive' encoding conditions. In fact, previous studies using the inducer-diagnostic task design suggested that inducer task instructions affected diagnostic task performance only when subjects assumed that the instructed rules will need to be implemented later on. Only then, the instructed rules seemed to be well prepared and maintained within working memory (Liefooghe, De Houwer, & Wenke, 2013; Wenke, Gaschler, Nattkemper, & Frensch, 2009). As a recent example, Meiran et al. (2015, Exp. 4) showed that instructions were especially well maintained across diagnostic task performance when subjects knew in advance that subsequent instructed task implementation had to rely on the instructed rules under conditions where only a few expected subsequent practice trials would be insufficient for compensatory trial-and-error learning. Another earlier study showed that instructed rules only affected diagnostic task performance when subjects assumed that these rules would be behaviorally implemented later on rather than merely recalled (Liefooghe, Wenke, & De Houwer, 2012). Especially the latter finding is reminiscent of conditions that produce dissociations between the ability to report instructed rules and the ability to behaviorally implement these rules (Duncan, Emslie, Williams, Johnson, & Freer, 1996; Duncan et al., 2008; Luria, 1973) – dissociations that seem to be well captured by the distinction between 'declarative' and 'procedural' sub-domains of working memory (Oberauer, 2009). However, despite these empirical and theoretical advances, it remains unclear whether these previous findings are limited to the maintenance of instructed rules across diagnostic task execution or if already the initial encoding of instructed rules depends on an active intention to apply them later on. In study 1 we examined if this might be the case by assessing a direct behavioral maker of instruction encoding processes.

Study 2 was designed to further elaborate on the results of study 1 and to assess the potential benefit of repeated instructions and to probe the involvement of working memory processes. The latter aim was inspired by recent theoretical considerations of instruction-based learning which have adopted the notion that instruction-based learning might primarily rely on working memory (WM) binding and maintenance processes (Liefooghe et al., 2013; Meiran & Cohen-Kdoshay, 2012; Meiran, Cole, & Braver, 2012; Ruge & Wolfensteller, 2010). More specifically, instruction-based learning can be conceptualized as the ad-hoc binding of distinct long-term memory (LTM) entries such as stimulus and response features of the instructed rule. Such ad-hoc bindings within working memory have been suggested to be established within a capacity-limited 'region of direct access' of LTM as opposed to 'activated LTM' which relies on permanent associations established through repeated exposure (Oberauer, 2009). The term 'binding' is used rather generically in this context but it is inspired by the theoretical notion that instructed S-R links might initially be stored in rather short-lived 'ad-hoc' working memory representations which contrasts with longer lasting representations in LTM often termed 'associations' (Henson, Eckstein, Waszak, Frings, & Horner, 2014; Liefooghe et al., 2013; Moeller & Frings, 2014). Notably, however, despite these theoretical presumptions, a recent study explicitly testing for a relationship between instruction-based learning efficiency and different WM scores failed to find clear-cut evidence (Meiran, Pereg, Givon, Danieli, & Shahar, 2016). The authors suspected that such a relationship might depend on a sufficiently high level of complexity or difficulty (i.e., ad-hoc binding demands) of the newly instructed task. In the present study we therefore instructed novel 4:4 S-R mappings instead of the 2:2 S-R mappings employed by Meiran et al. (2016). Moreover, we employed a simple-span WM test instead of a complex-span WM test which we hypothesized might be more sensitive to the relatively simple maintenance requirements in the type of instruction-based learning protocol we used. We therefore cautiously predicted that we might observe a significant relationship between inter-individual differences in WM-span and instruction-based learning due to these higher learning demands. This might be even more pronounced for less familiar stimulus material (here: non-words vs. words) which should be more resource demanding with respect to ad-hoc binding processes and therefore served as a proxy for increased WM load.

Study 2, together with study 3, additionally addressed the second major aim of this paper to not only assess efficiency of instruction-based learning processes in itself but also to directly compare it with trial-and-error learning. Specifically we aimed at substantiating the wide-held theoretical claim that instruction-based learning should be more efficient than trial-and-error learning simply because negative feedback trials, that is, failed attempts, can largely be avoided (Doll, Jacobs, Sanfey, & Frank, 2009; Noelle, 1997; Ruge & Wolfensteller, 2016b; Wolfensteller & Ruge, 2012). Efficiency was assessed via performance *accuracy* as an index of how well previously encoded S-R links could be correctly retrieved and implemented. Moreover, we assessed increasing performance *speed* of repeated *correct* S-R link implementations as an index of 'short-term task automatization' (Mohr et al., 2016).

2. Study 1

2.1. Aims

The key assumption of study 1 was that instruction encoding should be a time-consuming process. We devised an experimental design that measured instruction encoding time in terms of directly observable response times. To this end, in a given S-R learning block, the correct response R to a novel stimulus S was instructed by an additional response cue (RC) presented 150 ms after S onset and subjects were required to execute that response (see Fig. 1). Since S-R instruction encoding was hypothesized to interfere with preparing and executing the response to the RC, we expected a prolongation of response times (RTs). This encoding-related RT increase should be especially pronounced for the first instructed trial for each novel S-R link. Subsequently, encoding time should be decreasing with repeated instructions for the same S-R link. In study 1 we realized three instructed repetitions for each of four unique S-R links per learning block. Note that the 'first repetition' actually refers to the first implementation of specific S-R